water

MDPI

Article

# Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia

**Ayesha S Rahman and Ataur Rahman ***

School of Engineering, Western Sydney University, Sydney 00917K, New South Wales, Australia; ayesha.rahman@uws.edu.au

* Correspondence: a.rahman@uws.edu.au

Received: 6 February 2020; Accepted: 9 March 2020; Published: 12 March 2020


**Abstract:** This paper examines the applicability of principal component analysis (PCA) and cluster analysis in regional flood frequency analysis. A total of 88 sites in New South Wales, Australia are adopted. Quantile regression technique (QRT) is integrated with the PCA to estimate the flood quantiles. A total of eight catchment characteristics are selected as predictor variables. A leave-one-out validation is applied to determine the efficiency of the developed statistical models using an ensemble of evaluation diagnostics. It is found that the PCA with QRT model does not perform well, whereas cluster/group formed with smaller sized catchments performs better (with a median relative error values ranging from 22% to 37%) than other clusters/groups. No linkage is found between the degree of heterogeneity in the clusters/groups and precision of flood quantile prediction by the multiple linear regression technique.

**Keywords:** regional flood frequency analysis; design floods; principal component regression; cluster analysis; Australian rainfall and runoff

## 1. Introduction

Design flood estimation in poorly gauged or ungauged catchments is often a common problem in flood frequency analysis (FFA). To mitigate this problem, design floods can be estimated using regional flood frequency analysis (RFFA) [1–3]. There are essentially three steps in the RFFA: (i) hydrologically homogeneous region identification; (ii) appropriate regional flood estimation model development after identification of homogeneous region(s); and (iii) developed RFFA model validation [4–6]. Although traditionally geographical proximity and political boundaries are often adopted to identify homogeneous regions [7–11], regions formed in such manner often lack hydrological similarity [12–16].

Some researchers have also adopted climatic and catchment characteristics to form homogeneous regions by using multivariate statistical techniques such as cluster analysis [8,17–25]. Ward's method is the most common method for clustering as it can form regions of roughly equivalent size and this method is considered to be more suitable for regionalization of flood data [26]. Finally, many researchers adopted a region of influence (ROI) approach to circumvent complications related to fixed state boundaries [6,11–13,16,25,27–33]. The ROI is much more flexible than the fixed region approach and can be easily incorporated with a range of RFFA methods. ROI can also effectively reduce regional heterogeneity by sub-region formation within a large region. In the ROI approach, various ways can be adopted to form sub-regions such as by geographical distance [34,35] or distance in a multi-dimensional catchment attributes space [4,6,11,29].

Once a region is formed based on an appropriate condition [17,34–38], the design flood can be estimated either by the index flood method [15,39,40] or a regression-based approach [6,41]. The regression-based approach is used in numerous studies. Both the ordinary least square and the generalized least square regression methods are adopted to estimate the coefficients of the prediction equations in regression-based approaches [6,29,41–46]. For regression-based RFFA models, a linear relationship is generally assumed between the dependent variable (flood quantiles) and predictor variables (catchment and climatic characteristics). It is also assumed that the predictor variables are uncorrelated among each other; however, in many practical situations this assumption is not fully satisfied [47].

Principal component analysis (PCA) is a statistical technique capable of generating statistically uncorrelated principal components (PC) which are the linear amalgamation of the original variables (catchment and climatic characteristics). PCA has been used previously to delineate homogenous regions. For example, Burn [48], DeCoursey and Deal [49], Hawley and McCuen [50], Kar et al. [51], and Nathan and McMahon [22] adopted PCA to generate the PCs consisting of different physical, hydrological, and meteorological variables, and accordingly used them in principal component regression (PCR) to estimate flood quantiles. Choi et al. [52], Haque et al. [53], Haque et al. [54], and Koo et al. [55] used PCs in a PCR for water demand forecasting. Although PCA can produce uncorrelated PCs, one shortcoming of PCA is that due to the use of variance as an objective function, statistically independent structures are not always guaranteed in PCA [53].

There is a lack of research on delineation of homogeneous regions in RFFA. To fill this research gap, this study examines the formation of homogeneous regions using PCA and cluster analysis. This investigates the use of PCR method in RFFA with ROI and fixed region approaches. Multiple linear regression (MLR) models are developed for the regions generated by cluster analysis to estimate flood quantiles. A leave-one-out validation technique is adopted to assess the performance of the developed models.

## 2. Study Area

A very large catchment can have a significantly different flood frequency behavior compared to smaller sized catchments. Australian Rainfall and Runoff (ARR) [32,56] recommends an upper limit of 1000 km$^2$ for small to medium sized catchments [6], which appears to be rational to select candidate catchments for this study. From New South Wales (NSW), Australia, a total of 88 catchments are selected to carry out this study. These are natural catchments and free from any major storage and land use change. These selected catchments have catchment areas varying from 8 to 1010 km$^2$. The mean of catchment area is found to be 352 km$^2$ and median is found to be 260 km$^2$. It is recommended in Rahman et al. [57] to select catchments that have at least 20 years of flood data to develop the RFFA models in Australia. For this study, the selected catchments show a record length of annual maximum (AM) flow data in the range from 25 to 82 years (mean of 41.5 years and median 37 years). The catchments selected, vary from mountain to coastal region. The mean annual rainfall for the chosen catchments ranges from 625–1955 mm with a mean of 1000 mm and a median of 910 mm. Figure 1 shows the location of the selected 88 catchments.

A summary of descriptive statistics of selected catchment and climatic characteristics for the selected 88 sites is presented in Table 1. The design rainfall intensity of six-hour duration and two-year return period ($I_{62}$) at each catchment centroid is obtained from Australian Bureau of Meteorology website. The shape factor (*SF*) is defined as the shortest distance between a catchment's centroid and outlet divided by the square root of catchment area ($A$). Stream density (*sden*) is obtained as the sum of all the streamlines on a 1:100000 topographic map divided by catchment area ($A$). Mean annual rainfall (*MAR*) and mean annual evapotranspiration (*MAE*) data for each catchment are obtained from Australian Bureau of Meteorology website. The fraction forest (*forest*) is obtained as the total forested area shown on a 1:100000 topographic map divided by catchment area ($A$). The mainstream slope ($S1085$) is obtained as the difference in elevations at 10% and 85% of the mainstream length (measured

from the catchment outlet) divided by 0.75 of mainstream length. The PCs are extracted using these characteristics to develop the PCR models using quantile regression technique (QRT) and to form regions using cluster analysis.

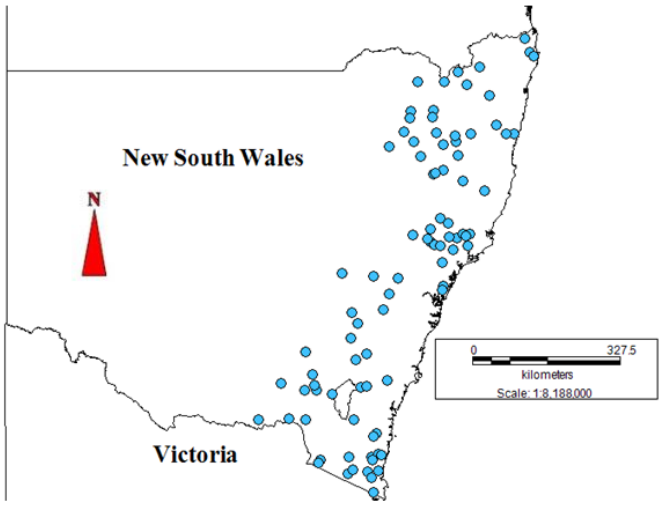


**Figure 1.** Spatial location of the selected 88 catchments in New South Wales (NSW), Australia.

**Table 1.** Summary of descriptive statistics of selected catchment and climatic characteristics.

| Variables | Range | Median | Mean | Standard Deviation |
|---|---|---|---|---|
| Catchment area ($A$) in km$^2$ | 8–1010 | 260 | 351.9 | 281.4 |
| Rainfall intensity ($I_{62}$) in mm/h | 31.3–87.3 | 43.1 | 45.4 | 11.3 |
| Shape factor ($SF$) | 0.3–1.6 | 0.8 | 0.8 | 0.2 |
| Stream density ($sden$) in /km | 0.5–5.5 | 2.8 | 2.7 | 1.1 |
| Mean annual rainfall ($MAR$) in mm | 626.2–1953.2 | 1000.3 | 909.9 | 304.5 |
| Mean annual evapo-transpiration ($MAE$) in mm/y | 980.4–1543.3 | 1223.7 | 1185.6 | 126.3 |
| Fraction forest ($forest$) | 0–1 | 0.5 | 0.5 | 0.3 |
| Mainstream slope ($S1085$) in m/km | 1.5–49.8 | 12.9 | 9.1 | 10.8 |

## 3. Methods

### *3.1. Principal Component Analysis*

Multiple linear regression (MLR) models get unstable with an increasing number of predictor variables, in particular, if these are highly correlated. PCA is one of the multivariate statistical techniques that can be used to deal with highly correlated variables in regression [58]. In PCA, original dataset of $n$ variables, which are correlated to various degrees are transformed to $n$ numbers of uncorrelated PCs. The PCs are linear transformation of the original variables in such a way that the original and the new variables have equal sums of the variances. Although the number of PCs and original variables are equal, the first few PCs explain the majority of the variance in the data set, reducing the dimensionality of the original data set [59]. The PCs are sequenced from the highest to the lowest variance, i.e., the first PC describes the data's highest variance proportion. The next highest variance is explained by the second PC and so on. The values of PCs can be obtained from Equations (1) and (2):

$$PC1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = \sum_{j=1}^{n} a_{1j}x_j, \quad (1)$$

$$PC2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = \sum_{j=1}^{n} a_{2j}x_j, \quad (2)$$

where $x_1, x_2, \ldots x_n$ are the original variables and $a_{jj}$ are the eigenvectors. The eigenvalues are the variances of the PCs. The covariance or correlation matrix of the data set is used to derive the coefficients $a_{jj}$, which are the eigenvectors. The eigenvalues of the data matrix can be calculated by Equation (3):

$$|C - \lambda I| = 0, \tag{3}$$

where $C$ is the correlation/covariance matrix, $\lambda$ is the eigenvalue, and $I$ is the identity matrix. The PC coefficients or the weights of the variables in the PCs are then calculated by Equation (4):

$$|C - \lambda I|a_{jj} = 0, \tag{4}$$

In the PCR analysis, PCs are used as predictor variable in MLR [60]. The general form of PCR model is as follows:

$$Y = \alpha + \beta_1 PC_1 + \beta_2 PC_2 + \cdots \beta_n PC_n, \tag{5}$$

where $Y$ is the dependent variable (which is flood quantile here), $\alpha$ is the model intercept, $\beta$'s are the regression coefficients.

### 3.2. Cluster Analysis

The statistical distance measurement representing the similarity (or dissimilarity) among the collections of attributes (similarity measurements) selected for each gauging site is used for grouping sites in cluster analysis. There are various clustering techniques available in the statistical literature [61] and are used to delineate hydrologically homogeneous regions [17,26,35,36,62–65]. Ward's method is the most commonly used method since this can produce clusters of similar sizes [26]. Hence, Ward's method is adopted in this study.

Ward's approach is an agglomerative hierarchical algorithm that starts with each site being its own cluster (or region). The algorithm successively merges clusters using a variance approach analysis in which the similitude between members in a region is measured in terms of the square error sum (*ESS*). For region $k$ containing $N_k$ sites, the *ESS* is calculated as:

$$ESS_k = \sum_{j=1}^{N} (x_j - \overline{x})^T (x_j - \overline{x}), \tag{6}$$

where $x_j = [x_1, x_2, \ldots, x_p]^T$ is a vector of $p$ characteristics measured at site $j$, and where each element denotes the mean value of a characteristic across the $N_k$ sites in the region. $ESS_k$ is calculated for the theoretical fusion of any two clusters at each step, and the actual fusions selected are those which minimize the increment in the total ESS across all regions.

### 3.3. Region of Influence Approach

Formation of regions without fixed boundaries was firstly carried out by Acreman [7]. Based on this, Burn 1990 a, b [12,13] introduced the ROI approach. In this approach, the individual site of interest (i.e., catchment where flood quantiles are to be estimated) forms its own region. Such identified regions can overlap and gauged sites for different sites of interest can be part of more than one ROI. The ROI may be formed for the site of interest using the group of sites in close proximity to the site of interest. More recently, the ROI approach has been adopted in ARR RFFA [32] and also in a study carried out by Rahman et al. [6]. A weighted Euclidean distance in an M-dimensional space may be used to measure the proximity. The distance metric can be defined by:

$$D_{i,j} = \left[ \sum_{m=1}^{M} W_m \left( X_i^m - X_j^m \right)^2 \right]^{1/2}, \tag{7}$$

where $D_{i,j}$ is the weighted Euclidean distance between site $i$ and $j$, $M$ is the number of features incorporated in the distance measure, and the $X$ terms represent standardized values for feature $m$ at

site $i$ and site $j$, and $W_m$ is a weight applied to attribute $m$, which reflects the relative significance of the feature. Standardization of attributes is performed to remove units and therefore *bias* due to scaling of the attributes can be avoided.

### 3.4. Homogeneity Assessment

Here, the Hosking and Wallis' [15] criteria of heterogeneity is adopted, which is based on L moments. A group of catchments is considered to be heterogeneous if H is too high. When H is smaller than 1, the group is taken as 'acceptably homogeneous', H falls between +1 to +2, the group is taken as 'possibly heterogeneous', and $H \geq 2$ it indicates a 'definitely heterogeneous' group. Furthermore, there are three different measures of H: $H_1$ is based on L coefficient of variation, $H_2$ is based on L coefficient of variation, and L coefficient of skewness and $H_3$ is based on L coefficient of skewness and L coefficient of kurtosis [15].

### 3.5. Evaluation Statistics

A leave-one-out validation technique is adopted for assessing the performance of the developed RFFA models. Based on leave-one-out validation, during the construction of a model, a site is left out in each phase, i.e., this site is treated as an ungauged site. The following performance statistics for each of the models are computed using predicted flood quantile ($Q_{pred}$) and observed flood quantile ($Q_{obs}$): relative error (*RE*), median absolute relative error (med_$RE_r$), $Q_{pred}/Q_{obs}$ ratio, median $Q_{pred}/Q_{obs}$ ratio (med_$Q_{pred}/Q_{obs}$), mean square error (*MSE*), root mean square error (*RMSE*), *bias* (*BIAS*), relative *bias* (*RBIAS*), relative root mean square error (*RRMSE*), and root mean square normalized error (*RMSNE*):

$$RE = \frac{Q_{pred} - Q_{obs}}{Q_{obs}} \times 100, \tag{8}$$

$$RE_r = median[abs(RE)], \tag{9}$$

$$MSE = mean[\left(Q_{pred} - Q_{obs}\right)^2], \tag{10}$$

$$RMSE = \sqrt{MSE}, \tag{11}$$

$$Bias = mean\left(Q_{pred} - Q_{obs}\right), \tag{12}$$

$$RBias = \left[mean\left(\frac{Q_{pred} - Q_{obs}}{Q_{obs}}\right)\right] \times 100, \tag{13}$$

$$RRMSE = \frac{\sqrt{mean[\left(Q_{pred} - Q_{obs}\right)^2]}}{mean(Q_{obs})}, \tag{14}$$

$$RMSNE = \sqrt{mean[\left(\frac{Q_{pred} - Q_{obs}}{Q_{obs}}\right)^2]}, \tag{15}$$

$Q_{obs}$ is the observed flood quantile at site $i$. $Q_{obs}$ is obtained by carrying out at-site FFA using LP3 distribution by FLIKE software [66]. In this study, six flood quantiles with annual exceedance probabilities (AEPs) of 50%, 20%, 10%, 5%, 2%, and 1% are considered as dependent variables.

## 4. Results and Discussion

### 4.1. Principal Component Analysis

Table 2 shows the magnitude and type of correlation between the original catchment and climatic characteristics (predictors). Correlation between the predictors, i.e., catchment area, rainfall intensity, shape factor, stream density, mean annual rainfall, mean annual evapo-transpiration, slope and fraction forest ('*A*', '$I_{62}$', '*SF*', '*sden*', '*MAR*', '*MAE*', '*S*1085', and '*forest*', respectively) are calculated using

the method described in methods section. Looking into Figure 2 and Table 2, it can be seen that catchment area has a negative correlation (−0.208; *p*-value 0.052) with the rainfall intensity indicating if the catchment area is large, rainfall intensity decreases. Rainfall intensity has a positive correlation with the variables shape factor, stream density, fraction forest, mean annual rainfall and mean annual evapo-transpiration, where the maximum positive correlation is between mean annual rainfall and rainfall intensity being 0.83 (*p*-value ≈ 0) and mean annual evapo-transpiration and rainfall intensity being 0.67 (*p*-value ≈ 0). This indicates that, if the rainfall intensity increases, mean annual rainfall also increases. Although these values are not close to ± 1, they have a very small *p* values (<0.10) indicating that these correlations are significant. Slope has a positive correlation of 0.387 with fraction forest and a negative correlation of −0.286 with mean annual evapo-transpiration where both the coefficients have smaller *p* values (<0.10). The variable fraction forest also has a positive correlation with the variable mean annual rainfall (0.405; *p*-value ≈ 0) and mean annual evapo-transpiration has positive correlations with stream density and mean annual rainfall (0.392 and 0.533; *p*-values ≈ 0). All the other correlation coefficients range from −0.007 to 0.303, which are statistically insignificant.

**Table 2.** Correlation coefficients with their corresponding *p*-values between the independent variables.

| | ***A*** | ***$I_{62}$*** | ***SF*** | ***sden*** | ***forest*** | ***MAR*** | ***MAE*** |
|---|---|---|---|---|---|---|---|
| $I_{62}$ | −0.208 | | | | | | |
| | 0.052 | | | | | | |
| *SF* | −0.054 | 0.035 | | | | | |
| | 0.619 | 0.746 | | | | | |
| *sden* | −0.175 | 0.367 | 0.037 | | | | |
| | 0.102 | 0 | 0.733 | | | | |
| *forest* | −0.116 | 0.33 | −0.007 | 0.046 | | | |
| | 0.283 | 0.002 | 0.951 | 0.667 | | | |
| *MAR* | −0.314 | 0.83 | −0.058 | 0.361 | 0.405 | | |
| | 0.003 | 0 | 0.592 | 0.001 | 0 | | |
| *MAE* | −0.094 | 0.671 | 0.136 | 0.392 | −0.031 | 0.533 | |
| | 0.381 | 0 | 0.206 | 0 | 0.771 | 0 | |
| *S*1085 | −0.331 | −0.121 | 0.051 | −0.081 | 0.387 | −0.021 | −0.286 |
| | 0.002 | 0.262 | 0.637 | 0.451 | 0 | 0.844 | 0.007 |

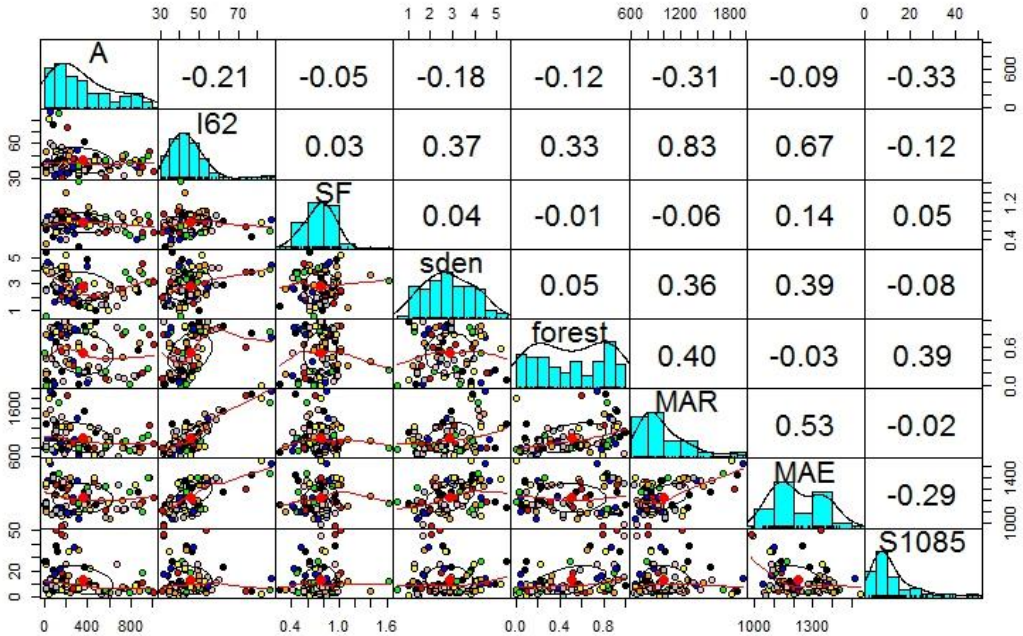


**Figure 2.** Correlation between the predictor variables before applying principal component analysis (PCA).

From the above discussion it is possible to say that some of the predictor variables have a notable degree of correlation between them. Therefore, PCA is applied to the eight selected predictors, i.e., catchment area, rainfall intensity, shape factor, stream density, mean annual rainfall, mean annual evapo-transpiration, slope and fraction forest ('*A*', '$I_{62}$', '*SF*', '*sden*', '*MAR*', '*MAE*', '*S*1085', and '*forest*', respectively) to achieve the uncorrelated eight PCs. Figure 3 shows the transformed PCs without any correlation. The eigenvalues with their percentage of contribution represent the quantity of variability in the data and they are presented in Table 3. Table 3 confirms that the first two PCs explain the maximum degree of variability of the data set with the proportion of *PC*1 and *PC*2 being 35.3% and 20.5%, respectively. The proportions of other PCs (PC3, PC4, PC5, PC6, PC7, and PC8) range 1.8%–13.4%. Although, *PC*1 and *PC*2 have the highest percentages among all the PCs, however, the cumulative of these two PCs only accounts for 55.8% variance, meaning these two PCs can only explain half of the variability in the dataset. To explain at least 85% of the variability in the data the first five PCs are required though the individual percentage is quite low for some PCs.

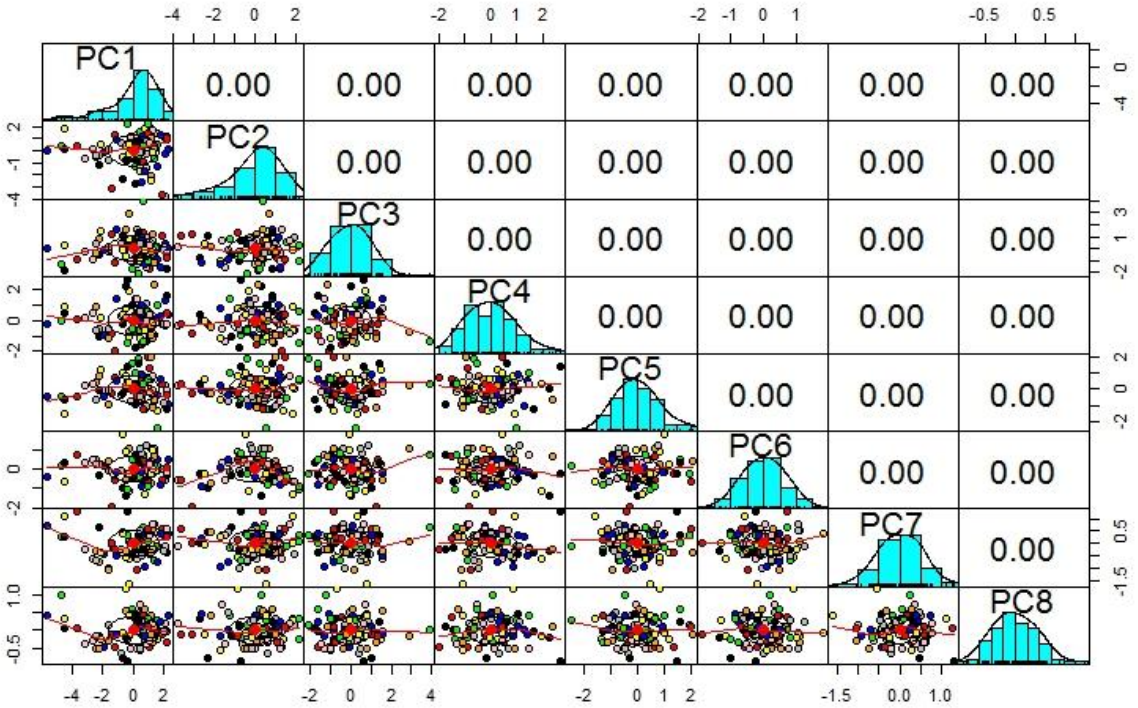


**Figure 3.** Uncorrelated principal components (PCs) after application of PCA.

**Table 3.** Eigenvalue of different components and their significance level.

| Name | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 |
|---|---|---|---|---|---|---|---|---|
| Eigenvalue | 2.822 | 1.641 | 1.070 | 0.915 | 0.702 | 0.432 | 0.278 | 0.141 |
| Proportion | 0.353 | 0.205 | 0.134 | 0.114 | 0.088 | 0.054 | 0.035 | 0.018 |
| Cumulative | 0.353 | 0.558 | 0.692 | 0.806 | 0.894 | 0.948 | 0.982 | 1 |

### *4.2. Development and Testing of Regression Equation in Fixed Region and ROI Framework*

To select the most impactful PCs, a significance level (*p*-value) for each of the PCs in the regression analysis is examined and the selection criterion is set to $p \leq 0.1$. Based on this criterion, *PC*1, *PC*2, PC3, PC4, and PC5 are found to be the significant ones to be used in the development of regression equation to estimate the flood quantiles. However, the coefficient of determination ($R^2$) and adjusted coefficient of determination ($adj\_R^2$) are found to be quite small for the regression equations, 0.46 and 0.43, respectively. The developed regression equation is then tested in both fixed region and ROI framework with leave-one-out validation. To apply fixed region (FR) approach, all the 88 sites are

grouped together as 'one region' and all the flood quantiles for each site for the six AEPs (50%, 20%, 10%, 5%, 2%, and 1%) are estimated by leave-one-out validation.

In the ROI framework, at first, 10 sites are grouped together to form one region and the flood quantiles are estimated. Afterwards, at each of the iterations, five new sites are added to form a larger region until the site number reached 30. When the site number reached 30, ten sites are added at each of the iterations to form a larger region until the number of sites reached 80. Leave-one-out is applied at each of the iterations for validation.

Table 4 shows the statistical evaluation for 5% AEP flood. The first column shows the statistical evaluations that are calculated for both the fixed region and ROI cases, which are *MSE*, *RMSE*, *BIAS*, *RBIAS*, *RRMSE*, *RMSNE*, $R_r$, and med_$Q_{pred}/Q_{obs}$ based on observed and predicted values for 5% AEP flood. Looking into Table 4, it is found that for 5% AEP flood, fixed region has the lowest *MSE*. The lowest $R_r$ is found in case of KNN80 (46.59%) and med_$Q_{pred}/Q_{obs}$ close to 1 is found in case of KNN15. From Table 4, it is clear that KNN10 performs the poorest with largest *MSE* and $R_r$, which means it is preferable to select more than ten sites to form a region to use PCR. Durocher et al. [67] carried out a study in Southern Quebec (Canada) and their results show a *RMSE* in the range of 38 $m^3/s$ and 45 $m^3/s$ in case of 10% and 1% AEP floods using spatial copula method. For the same dataset in Québec a number of studies [47,68,69] were carried out. The results show *RMSE* values being 41 $m^3/s$ to 51 $m^3/s$ for 10% AEP flood, and 49 $m^3/s$ to 70 $m^3/s$ for 1% AEP flood. These studies were carried out using ordinary kriging in PCA-space, generalized additive model and single artificial neural network. Studies carried out by Durocher et al. [67], Chokmani and Ouarda [68], Chebana et al. [20] and Shu and Ouarda [69] show *RBIAS* values ranging from −5% to −20% for 10% AEP flood and −7% to −27% for 1% AEP flood. A study carried out by Rahman et al. [6] found *RBIAS* values ranging from 22% to 69% for the six AEP floods.

For further clarification on the number of sites required to form regions in case of using this technique in FFA, boxplots are examined based on their *RE* and $Q_{pred}/Q_{obs}$ ratio values for both the fixed region and ROI framework.

Figures 4 and 5 show boxplots of the *RE* and $Q_{pred}/Q_{obs}$ ratio values for both the fixed region and ROI framework for 5% AEP flood. Both Figures 4 and 5 starts with fixed region approach and have all the ROI approaches presented one by one after the fixed region. Looking at Table 3, one can see that fixed region performs better than the rest of the ROI models. However, Figures 4 and 5 show that, although the box size is smaller (i.e., a smaller error range), the median line is not close to the expected line (expected lines are set at zero and one for Figures 4 and 5, respectively) for the fixed region. KNN10 shows a similar performance as presented in Table 4. KNN15 and KNN25 both show promising results in Figures 4 and 5 with a smaller box size and median value being very close to the median line. However, it seems that KNN25 has smaller error bars than KNN15. There are number of outliers for all the models as shown by small circles, but they are not all visible in the figures as the range for the boxplots are set in the range of −300 to +300 for the *RE* and −2 to +3 for $Q_{pred}/Q_{obs}$ ratio values to have a greater visibility. In Figure 5, it is seen that none of the top error bars are visible in the set range bringing in the question of how well they fit the regression analysis. The rest of the ROI models show that they represent a poorer fit with bigger box size and median *RE* being far away from the expected line.

**Table 4.** Statistical evaluation for fixed region (FR) and region of influence (ROI) for 5% annual exceedance probability (AEP) flood.

| 5% AEP | KNN10 | KNN15 | KNN20 | KNN25 | KNN30 | KNN40 | KNN50 | KNN60 | KNN70 | KNN80 | FR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSE | 443296.69 | 228884.66 | 183799.06 | 187180.23 | 179135.55 | 181519.00 | 170279.10 | 166945.45 | 163644.38 | 163850.83 | 163447.17 |
| RMSE | 665.81 | 478.42 | 428.72 | 432.64 | 423.24 | 426.05 | 412.65 | 408.59 | 404.53 | 404.78 | 404.29 |
| BIAS | −65.51 | 1.20 | −11.94 | −13.12 | −0.82 | −20.07 | −14.73 | −21.27 | −18.68 | −6.20 | −0.29 |
| RBIAS | 22.24 | −0.44 | 55.31 | 63.94 | 54.40 | 61.69 | 65.54 | 56.98 | 54.34 | 69.90 | 65.48 |
| RRMSE | 0.11 | 0.00 | 0.02 | 0.02 | 0.00 | 0.03 | 0.03 | 0.04 | 0.03 | 0.01 | 0.00 |
| RMSNE | 5.40 | 3.15 | 3.28 | 3.03 | 2.38 | 2.77 | 2.28 | 2.05 | 2.05 | 2.69 | 2.43 |
| med_$R_r$ | 59.03 | 53.41 | 54.46 | 52.61 | 55.12 | 51.48 | 49.01 | 50.74 | 47.15 | 46.59 | 48.08 |
| med_$Q_{pred}/Q_{obs}$ | 1.03 | 1.01 | 1.16 | 1.07 | 1.14 | 1.19 | 1.20 | 1.16 | 1.16 | 1.18 | 1.17 |

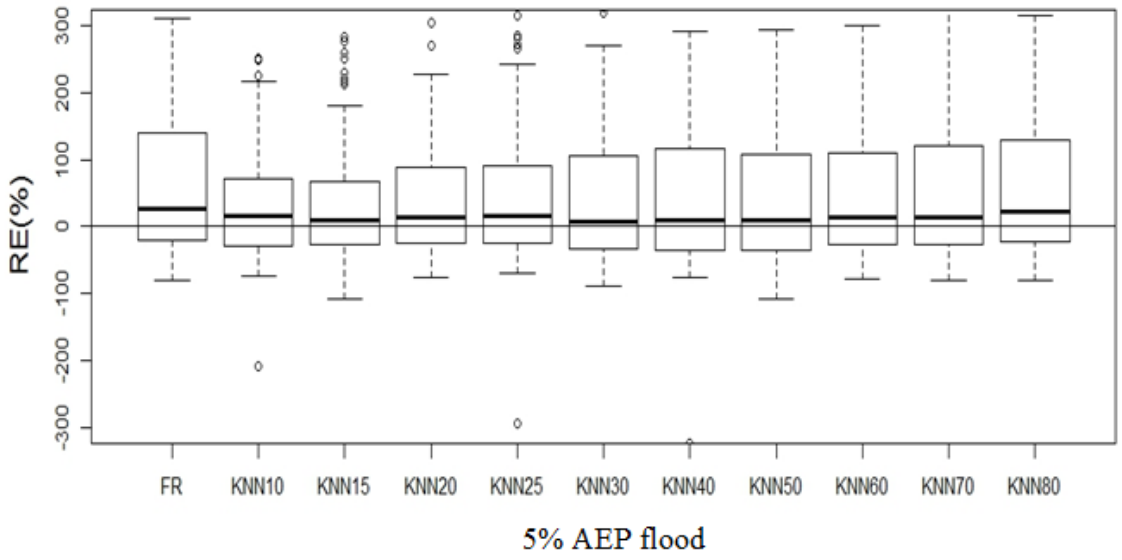


**Figure 4.** Boxplots for *RE* (%) of fixed region (FR) and ROI (in case of 5% AEP flood).

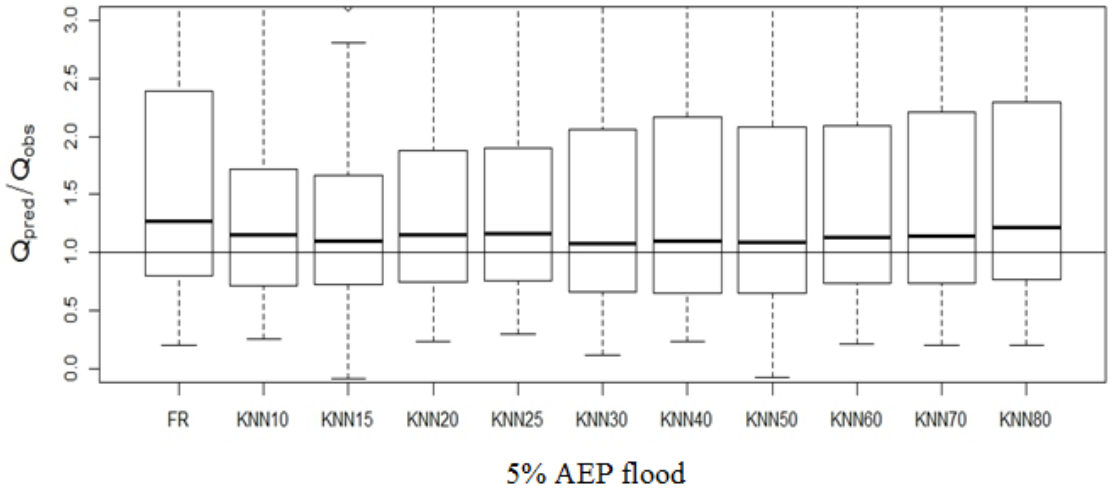


**Figure 5.** Boxplots for $Q_{pred}/Q_{obs}$ of fixed region (FR) and ROI (in case of 5% AEP flood).

Tables 5 and 6 compare the *RE* and the $Q_{pred}/Q_{obs}$ ratio values for all the AEPs for both the fixed region and ROI, respectively. Rows 2 to 18 of Tables 5 and 6 show the mean, median and standard deviation (Std_Dev) of the selected AEPs for both the fixed region and ROI models, respectively. The last three rows show the overall mean, median, and Std_Dev for the AEPs. All the *RE* values are transformed to their absolute values by ignoring their sign. All the lowest values in case of both $Q_{pred}/Q_{obs}$ ratio values and *RE* values are presented with blue color in both Tables 5 and 6. Although KNN25 comes out as the best model out of all of them leaving KNN15 behind, however from Tables 5 and 6, it seems that KNN15 outperforms KNN25 especially in the case of $Q_{pred}/Q_{obs}$ ratio values. A fixed region approach or KNN80 does not show any better performance in this case. As seen earlier, KNN10 shows the worst results. The other models generate a mixture of results. In some cases, a very large *RE* (%) and $Q_{pred}/Q_{obs}$ ratios are also found (i.e., for stations 206026, 210068, 210076, and 222016). As seen from the $R^2$ and $adj\_R^2$ values, this regression model is found to be representing a poor fit. The analysis for both the fixed region and ROI framework also supports this finding.

**Table 5.** Mean, median and standard deviation of relative error for fixed region (FR) and ROI.

| RE | | FR | KNN10 | KNN15 | KNN20 | KNN25 | KNN30 | KNN40 | KNN50 | KNN60 | KNN70 | KNN80 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50% | Mean_abs | 139.75 | 229.93 | 114.81 | 152.82 | 123.56 | 133.99 | 122.83 | 122.97 | 127.89 | 131.03 | 133.88 |
| | Median_abs | 51.76 | 63.84 | 42.70 | 45.76 | 42.70 | 48.91 | 42.19 | 44.84 | 46.70 | 49.10 | 50.12 |
| | Std Dev_abs | 356.39 | 623.98 | 156.38 | 310.14 | 204.57 | 238.60 | 218.48 | 247.15 | 227.90 | 266.56 | 323.67 |
| 20% | Mean_abs | 122.18 | 206.85 | 117.88 | 134.93 | 132.02 | 126.40 | 117.35 | 112.29 | 111.24 | 113.35 | 125.28 |
| | Median_abs | 48.32 | 51.22 | 51.25 | 54.02 | 43.64 | 50.44 | 46.03 | 50.88 | 50.09 | 46.27 | 45.36 |
| | Std Dev_abs | 276.38 | 530.31 | 164.84 | 242.12 | 211.61 | 243.99 | 264.68 | 230.54 | 244.19 | 246.78 | 286.58 |
| 10% | Mean_abs | 118.13 | 208.47 | 130.45 | 140.34 | 135.02 | 127.57 | 125.56 | 115.12 | 105.84 | 107.25 | 122.45 |
| | Median_abs | 48.62 | 55.19 | 55.39 | 54.23 | 48.77 | 50.32 | 49.57 | 51.92 | 53.24 | 47.21 | 41.70 |
| | Std Dev_abs | 223.95 | 504.74 | 182.24 | 260.47 | 236.11 | 225.52 | 247.62 | 200.09 | 197.48 | 200.13 | 247.30 |
| 5% | Mean_abs | 118.95 | 213.73 | 153.93 | 143.57 | 135.83 | 127.49 | 132.75 | 121.48 | 111.44 | 108.40 | 124.36 |
| | Median_abs | 48.09 | 59.03 | 53.41 | 54.46 | 52.62 | 55.13 | 51.48 | 49.02 | 50.74 | 47.15 | 46.59 |
| | Std Dev_abs | 213.32 | 498.85 | 276.47 | 297.00 | 272.61 | 201.63 | 244.73 | 193.92 | 172.80 | 174.95 | 239.47 |
| 2% | Mean_abs | 130.02 | 225.32 | 197.48 | 154.64 | 144.06 | 140.61 | 150.67 | 137.73 | 126.73 | 118.43 | 134.54 |
| | Median_abs | 52.16 | 66.95 | 58.87 | 56.14 | 58.25 | 60.52 | 54.35 | 52.14 | 50.76 | 53.10 | 53.28 |
| | Std Dev_abs | 258.97 | 516.14 | 542.05 | 374.04 | 359.08 | 251.26 | 308.61 | 256.72 | 218.79 | 194.39 | 279.09 |
| 1% | Mean_abs | 143.45 | 239.70 | 248.66 | 174.50 | 164.45 | 163.20 | 170.96 | 156.71 | 144.20 | 130.90 | 147.16 |
| | Median_abs | 53.56 | 71.95 | 68.67 | 61.21 | 59.66 | 59.80 | 52.97 | 50.99 | 49.75 | 47.46 | 51.92 |
| | Std Dev_abs | 322.07 | 548.06 | 848.84 | 458.31 | 459.05 | 351.10 | 400.20 | 341.21 | 294.90 | 239.05 | 335.53 |
| Overall mean | | 128.75 | 220.67 | 160.53 | 150.13 | 139.16 | 136.54 | 136.69 | 149.15 | 121.22 | 131.30 | 133.48 |
| Overall median | | 49.77 | 61.39 | 54.90 | 54.42 | 51.31 | 55.06 | 50.63 | 47.42 | 50.17 | 45.15 | 45.95 |
| Overall Std Dev | | 278.68 | 536.23 | 443.19 | 330.54 | 303.33 | 255.48 | 286.29 | 350.60 | 228.51 | 255.40 | 264.04 |

**Table 6.** Mean, median and standard deviation of $Q_{pred}/Q_{obs}$ ratio for fixed region (FR) and ROI.

| Ratio | | FR | KNN10 | KNN15 | KNN20 | KNN25 | KNN30 | KNN40 | KNN50 | KNN60 | KNN70 | KNN80 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50% | Mean_abs | 2.00 | 2.72 | 1.67 | 2.09 | 1.84 | 1.96 | 1.81 | 1.84 | 1.86 | 1.87 | 1.94 |
| | Median_abs | 1.16 | 1.09 | 1.10 | 1.12 | 1.15 | 1.12 | 1.10 | 1.12 | 1.17 | 1.16 | 1.14 |
| | Std Dev_abs | 3.44 | 6.20 | 1.60 | 2.95 | 2.02 | 2.39 | 2.10 | 2.33 | 2.12 | 2.56 | 3.12 |
| 20% | Mean_abs | 1.85 | 2.53 | 1.69 | 1.90 | 1.90 | 1.85 | 1.79 | 1.77 | 1.71 | 1.73 | 1.89 |
| | Median_abs | 1.14 | 1.14 | 1.09 | 1.19 | 1.20 | 1.23 | 1.09 | 1.18 | 1.09 | 1.11 | 1.16 |
| | Std Dev_abs | 2.68 | 5.23 | 1.60 | 2.38 | 2.08 | 2.36 | 2.57 | 2.20 | 2.36 | 2.40 | 2.80 |
| 10% | Mean_abs | 1.83 | 2.53 | 1.81 | 1.93 | 1.92 | 1.87 | 1.88 | 1.81 | 1.72 | 1.73 | 1.89 |
| | Median_abs | 1.15 | 1.17 | 1.24 | 1.19 | 1.13 | 1.33 | 1.12 | 1.22 | 1.17 | 1.14 | 1.13 |
| | Std Dev_abs | 2.23 | 4.98 | 1.70 | 2.62 | 2.38 | 2.22 | 2.44 | 1.97 | 1.94 | 1.98 | 2.46 |
| 5% | Mean_abs | 1.88 | 2.58 | 2.04 | 1.97 | 1.94 | 1.87 | 1.98 | 1.90 | 1.80 | 1.76 | 1.93 |
| | Median_abs | 1.19 | 1.26 | 1.19 | 1.24 | 1.12 | 1.21 | 1.22 | 1.22 | 1.21 | 1.18 | 1.20 |
| | Std Dev_abs | 2.18 | 4.93 | 2.61 | 3.02 | 2.79 | 2.07 | 2.47 | 1.99 | 1.77 | 1.79 | 2.44 |
| 2% | Mean_abs | 2.00 | 2.67 | 2.45 | 2.08 | 2.07 | 2.04 | 2.16 | 2.07 | 1.96 | 1.86 | 2.04 |
| | Median_abs | 1.17 | 1.28 | 1.20 | 1.13 | 1.21 | 1.23 | 1.22 | 1.21 | 1.26 | 1.21 | 1.21 |
| | Std Dev_abs | 2.70 | 5.12 | 5.29 | 3.83 | 3.68 | 2.61 | 3.16 | 2.68 | 2.30 | 2.07 | 2.89 |
| 1% | Mean_abs | 2.15 | 2.77 | 2.95 | 2.29 | 2.27 | 2.27 | 2.37 | 2.25 | 2.14 | 1.99 | 2.17 |
| | Median_abs | 1.27 | 1.22 | 1.29 | 1.22 | 1.33 | 1.33 | 1.17 | 1.20 | 1.26 | 1.23 | 1.26 |
| | Std Dev_abs | 3.33 | 5.47 | 8.36 | 4.68 | 4.69 | 3.62 | 4.09 | 3.54 | 3.08 | 2.54 | 3.47 |
| Overall mean | | 1.95 | 2.63 | 2.10 | 2.04 | 1.99 | 1.98 | 2.00 | 1.94 | 1.87 | 1.82 | 1.98 |
| Overall median | | 1.17 | 1.18 | 1.14 | 1.17 | 1.19 | 1.19 | 1.16 | 1.18 | 1.18 | 1.17 | 1.20 |
| Overall Std Dev | | 2.79 | 5.31 | 4.34 | 3.32 | 3.08 | 2.59 | 2.87 | 2.50 | 2.29 | 2.23 | 2.87 |

### *4.3. Application of Cluster Analysis*

#### 4.3.1. Cluster Formation

Figure 6 shows the dendrogram by hierarchical cluster analysis (Ward's method) using all the predictors (standardized to 0 mean and unit variance). A total of five clusters are identified in Figure 6; each of the clusters has more than seven stations (cluster 4 has 23 stations, whereas the other clusters have 15 to 18 stations). The details of the five clusters are provided in Table 7. The median of streamflow record lengths for all the clusters are in the range 36 to 40 years. Cluster 5 contains catchments that are relatively large (area ranging from 454–1010 km$^2$ with a median value of 835.5 km$^2$). The median values of area for the other clusters are in the range 156–365 km$^2$. Looking into design rainfall intensity i.e., $I_{62}$, the median values range 38 mm–59 mm for all the clusters. The highest rainfall intensity is found in case of cluster 3, in the range 76 mm–133 mm. The variable '*SF*' is found to be similar for all the five clusters; whereas '*sden*' is found to be higher for clusters 1 and 3 and minimum for cluster 4. The variables '*MAR*' and '*MAE*' are found to be higher for cluster 1, i.e., 1480.2 mm and 1382.7 mm (median), respectively. Cluster 2 shows a relatively higher slope. Finally, fraction forest area is found to be relatively higher for clusters 1 and 2.

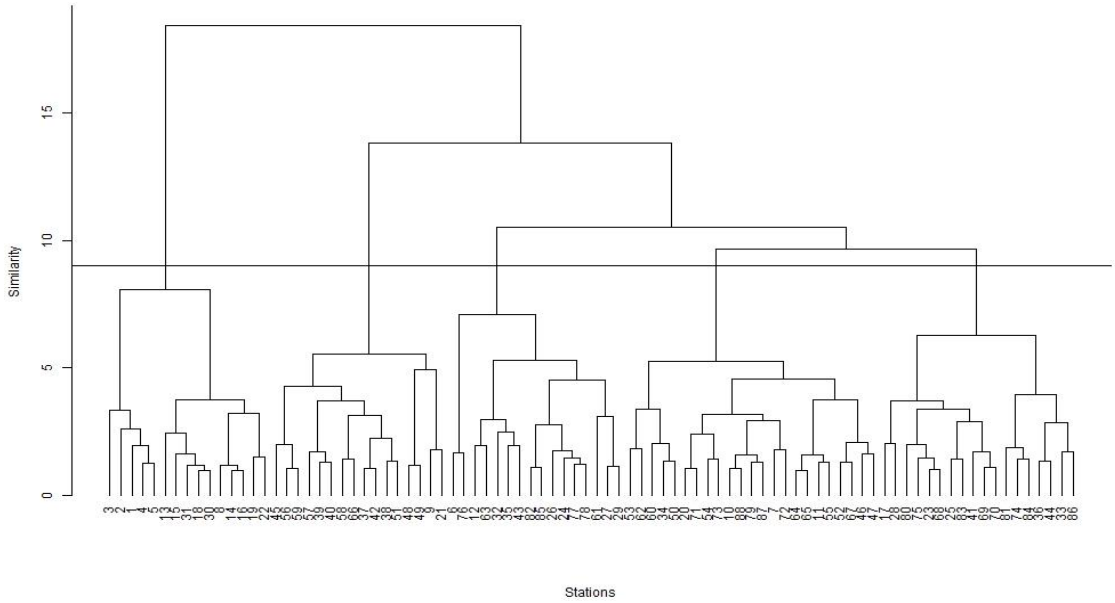


**Figure 6.** Result of Ward's hierarchical clustering method by using eight selected catchment characteristics (dendrogram). Figures 7 and 8 show the boxplots representing the distribution of the eight predictors for the clusters. Figure 7 shows the boxplots for '*A*', '$I_{62}$', '*SF*', and '*sden*' and Figure 8 shows the boxplots for '*MAR*', '*MAE*', '*S1085*', and '*forest*'. From Figures 7 and 8, it can be said that the stations in cluster 1 have the smallest catchment area range with the largest design rainfall intensity, as well as larger '*MAR*' and '*MAE*' values. These stations seem to have higher percentages of forest areas as well as along with higher stream density. Stations in cluster 2 have moderate catchment areas, along with '$I_{62}$', '*SF*', '*sden*', and '*MAR*'. However, the '*MAE*' is the highest in case of cluster 2. Cluster 3 seems to have all the predictors relatively uniformly distributed. Cluster 4 is characterized by medium catchment area, smaller '$I_{62}$', '*SF*', '*sden*', '*MAR*', '*MAE*', '*S1085*', and '*forest*'. The largest area is found for cluster 5, although the other predictors are quite small with large variance in fraction forest area.

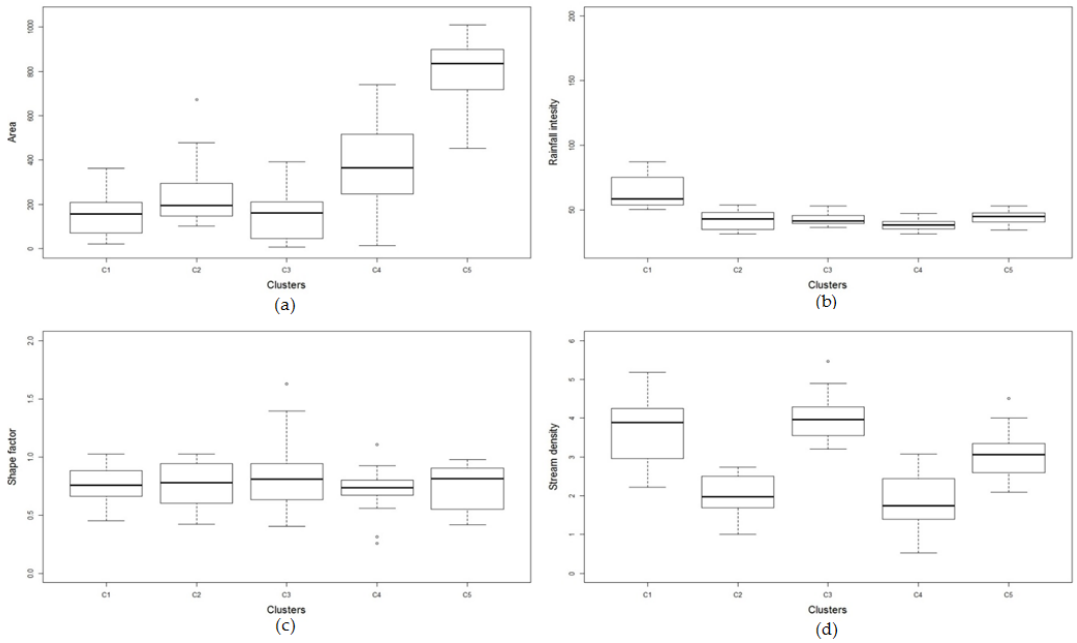


**Figure 7.** Boxplots of area, rainfall intensity, shape factor and stream density for all clusters: (**a**) boxplot of area; (**b**) boxplot of rainfall intensity; (**c**) boxplot of shape factor; and (**d**) boxplot of stream density.

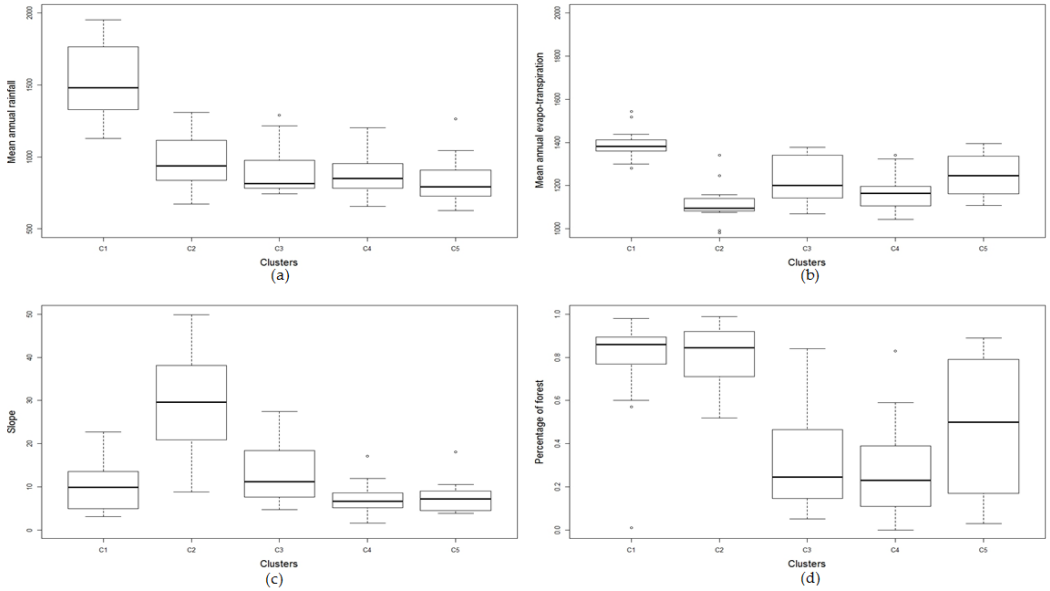


**Figure 8.** Boxplots of mean annual rainfall, mean annual evapo-transpiration, slope and fraction forest for all the clusters: (**a**) boxplot of mean annual rainfall; (**b**) boxplot of mean annual evapo-transpiration; (**c**) boxplot of slope; and (**d**) boxplot of fraction forest.

### 4.3.2. Homogeneity Analysis of the Clusters

To investigate the degree of homogeneity of the five clusters, the heterogeneity measure proposed by Hosking and Wallis [15] is applied to each cluster individually. According to Hosking and Wallis [38], any station showing $D_i \geq 3$ is considered to be discordant. Based on this criterion, no discordant station is found for clusters 1, 2, 4, and 5, yet one discordant station ($D_i = 3.01$) is found for cluster 3. The heterogeneity measure is applied to the five clusters to calculate $H$-statistics ($H_1$, $H_2$, and $H_3$). For cluster 3, although the $D_i$ value is not very large, the heterogeneity measure is applied twice; firstly, with all the discordant station in the cluster and secondly, removing the discordant station.

**Table 7.** Detail of the clusters formed by Ward's hierarchical clustering method.

| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| No. of stations | 15 | 16 | 16 | 23 | 18 |
| Period of records | 29–80 | 26–71 | 30–82 | 25–70 | 32–56 |
| (median) | (36) | (40) | (36) | (37) | (37.5) |
| Area ($km^2$) | 20–363 | 103–673 | 8–391 | 14–740 | 454–1010 |
| (median) | (156) | (194.5) | (161) | (365) | (835.5) |
| $I_{62}$ (mm) | 50–88 | 31–54 | 76–133 | 31–48 | 34–54 |
| (median) | (58.4) | (43.1) | (41.3) | (38.4) | (44.9) |
| SF | 0.4–1.02 | 0.4–1.02 | 0.4–1.7 | 0.2–1.2 | 0.4–1 |
| (median) | (0.8) | (0.8) | (0.8) | (0.7) | (0.8) |
| sden | 2.2–5.2 | 1–3 | 3.2–5.5 | 0.5–3.1 | 2–5 |
| (median) | (3.9) | (1.9) | (3.9) | (1.7) | (3.1) |
| MAR (mm) | 1128–1954 | 672–1310 | 744–1289 | 656–1204 | 626–1265 |
| (median) | (1480.2) | (937.5) | (815.2) | (851.2) | (791.9) |
| MAE (mm) | 1280–1544 | 980–1341 | 1069–1378 | 1044–1342 | 1107–1396 |
| (median) | (1382.7) | (1094.6) | (1200.5) | (1165.2) | (1245.9) |
| S1085 | 3–23 | 8–50 | 4–28 | 1–18 | 3–19 |
| (median) | (9.9) | (29.6) | (11.2) | (6.7) | (7.2) |
| forest | −1 | 0.5–1 | 0.05–1 | 0–0.83 | 0.03–1 |
| (median) | (0.9) | (0.9) | (0.3) | (0.2) | (0.5) |

Table 8 presents the heterogeneity measures for each cluster. It is visible from Table 8 that none of the clusters form homogeneous region. The lowest $H$-statistics is found for cluster 4; however, as the range is between $1 \leq H \leq 2$, cluster 4 is 'possibly heterogeneous'. Cluster 3 shows two $H$-statistics as one discordant station is found for cluster 3 (station 419029). Removal of the discordant station does not improve the result for cluster 3. Although the values of $H_2$ and $H_3$ for some clusters are smaller, $H_1$ is mostly indicative of the heterogeneity in the group, which is much higher than 1.00. It is of interest to check how these heterogeneous clusters perform in regional flood estimation. Hence, QRT is applied to each cluster in the next section with leave-one-out validation to validate the QRT.

**Table 8.** Heterogeneity measures for each cluster.

| | Number of Stations | $H_1$ | $H_2$ | $H_3$ |
|---|---|---|---|---|
| Cluster 1 | 15 | 5.11 | 4.93 | 3.71 |
| Cluster 2 | 16 | 7.38 | 2.92 | −0.05 |
| Cluster 3 | 16 | 7.59 | 6.13 | 3.62 |
| | 15 | 7.55 | 5.26 | 2.74 |
| Cluster 4 | 23 | 1.93 | 1.16 | 0.64 |
| Cluster 5 | 18 | 5.54 | 4.06 | 2.70 |

#### 4.3.3. Development of Prediction Equation and Performance Testing

For the development of prediction equation, the dependent (flood quantiles) and predictor variables are natural-log transformed (i.e., a log-log modelling is adopted). A stepwise procedure based on their level of significance ($p \leq 0.1$) is applied to select the best set of predictor variables. For different clusters, selection of the predictor variables generated different sets of equations because of their different levels of significance. The general regression equation for all clusters is given below for 5% AEP flood ($Q_{20}$) and the regression coefficients for each variable for each cluster are given in Table 9.

**Table 9.** Coefficients of each variable for each cluster in the regression (Equation (16)) for design flood estimation.

| Cluster | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.29 | 0.94 | 1.53 | 0 | 0.75 | 0 | −2.14 | −0.15 | 0 |
| 2 | −9.84 | 0.29 | 7.26 | −1.80 | −0.52 | −3.93 | 3.75 | −0.80 | 0 |
| 3 | −4.62 | 0.81 | 3.19 | −1.21 | 0 | 0 | 0 | 0 | 0 |
| 4 | −0.94 | 0.72 | 0.96 | 0 | 0.52 | 0 | 0 | 0 | 0 |
| 5 | 4.10 | 0.49 | 4.42 | −0.34 | 0.74 | −0.62 | −2.65 | 0 | −0.22 |

The general form of the regression equation for the clusters:

$$\ln Q_{20} = \beta_0 + \beta_1(\ln(A)) + \beta_2(\ln(I_{62})) + \beta_3(\ln(SF)) + \beta_4(\ln(sden)) + \beta_5(\ln(MAR)) + \beta_6(\ln(MAE)) + \beta_7(\ln(forest)) + \beta_8(\ln(S1085)), \quad (16)$$

The $R^2$ and $adj\_R^2$ for each model for all the clusters are found to be quite high except for cluster 5. In case of cluster 1, $R^2$ and $adj\_R^2$ values are 0.93 and 0.89, respectively, for the selected model; for cluster 2 they are 0.98 and 0.96, respectively; for cluster 3 they are 0.82 and 0.77, respectively; for cluster 4 these are 0.68 and 0.63, respectively, and for cluster 5 these are 0.66 and 0.48, respectively. It is evident that except cluster 5 the other four clusters generate regression models with satisfactory goodness-of-fit.

Figures 9 and 10 show the standardized residuals versus the fitted or predicted value plots and normality plots for 5% AEP flood for all the five clusters. It is necessary for the residuals of a linear regression model to satisfy homoscedastic pattern as heteroscedasticity in the residuals indicates that a non-linear model is more appropriate for the data. It is evident from the standardized residual versus the fitted value plots that the residuals lie between −2 and +2 and no specific pattern is visible in the plots. No pattern in the spread of the residuals indicates that the residuals are homoscedastic, which satisfies the linearity model assumption. The Q-Q plots do not completely follow the reference line except for cluster 2, yet there is no specific pattern, which indicates that the normality assumption is not grossly violated.

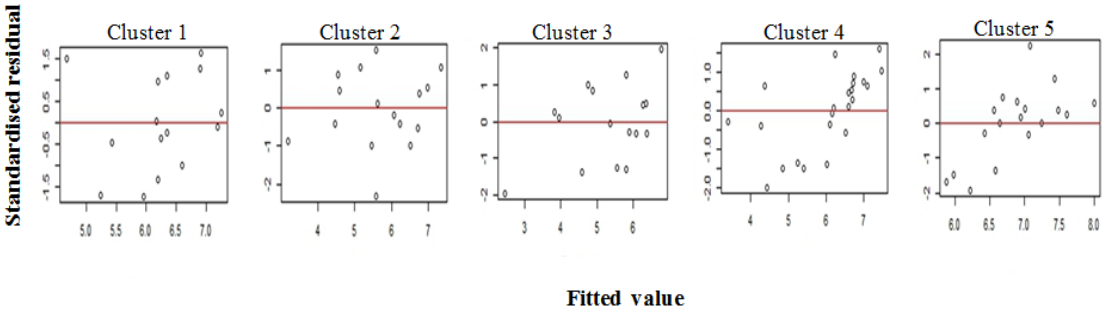


**Figure 9.** Standardized residuals vs fitted value plots for all clusters in case of 5% AEP flood.

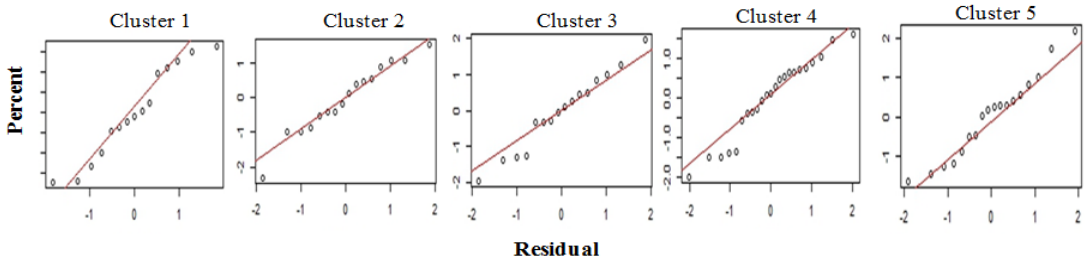


**Figure 10.** Normality plots for all clusters in case of 5% AEP flood.

Figures 11 and 12 show the boxplots of the selected quantiles for all the clusters in terms of *RE* and $Q_{pred}/Q_{obs}$ ratio values. The expected line in Figure 10 is set at zero as it indicates an un-biased model. For Figure 12, the expected line is set at one as the ratio being one is indicative of an unbiased model. Figure 11 has a set boundary of −300 to +300, whereas Figure 12 has a set boundary of −3 to 3 to have better visibility.

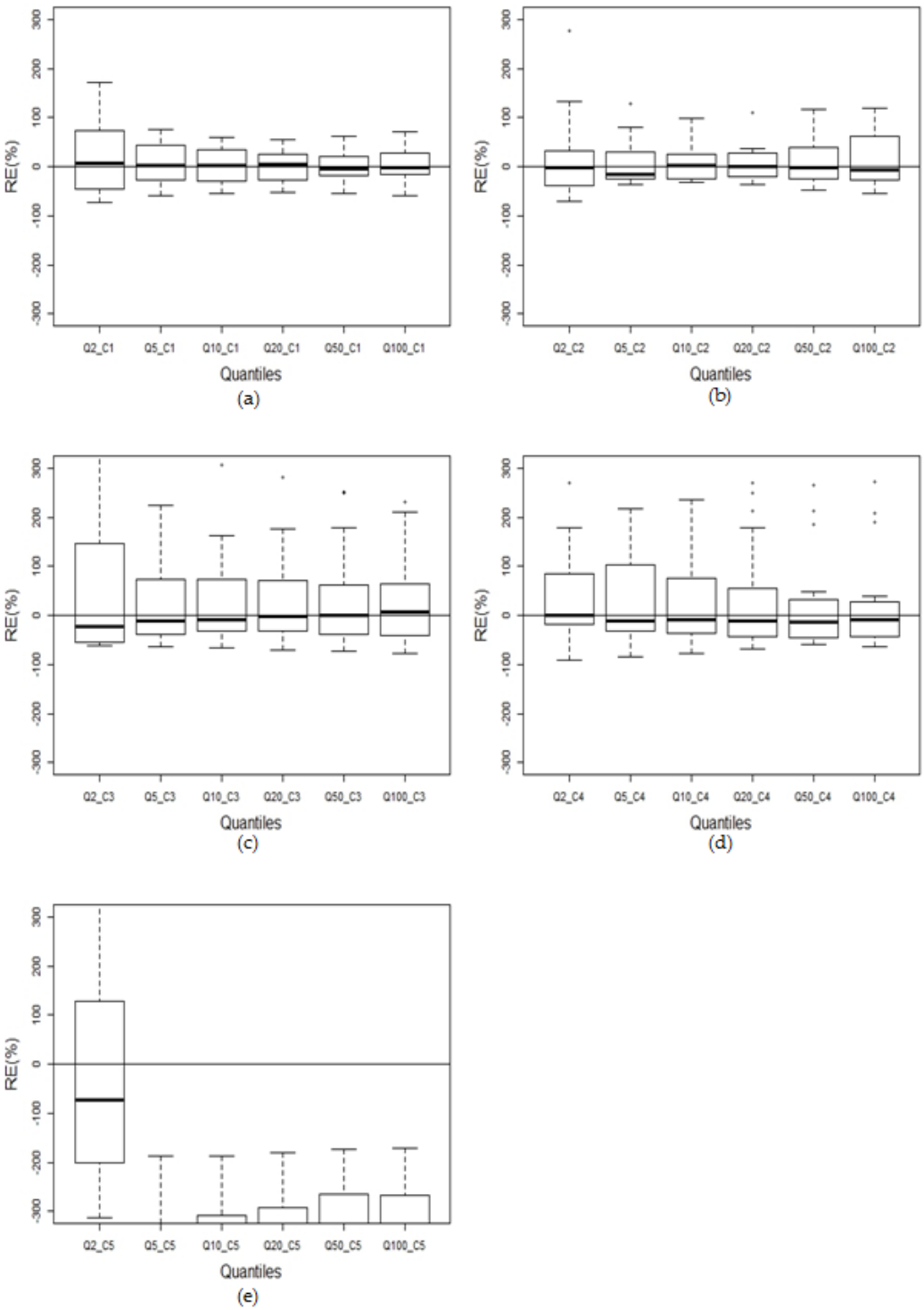


**Figure 11.** Boxplots of *RE* for all clusters and AEPs: (**a**) cluster 1 (C1); (**b**) cluster 2 (C2); (**c**) cluster 3 (C3); (**d**) cluster 4 (C4); and (**e**) cluster 5 (C5).

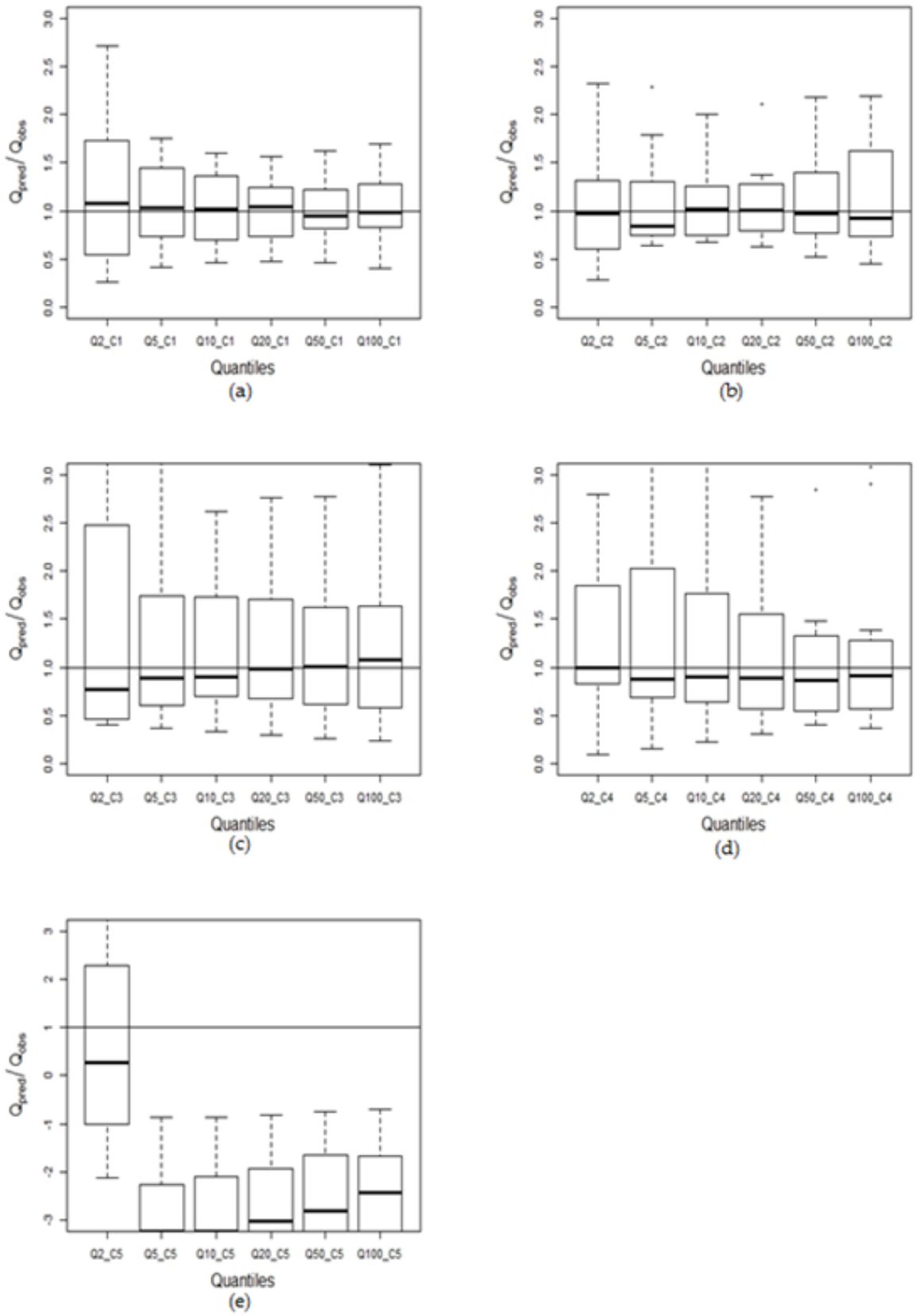


**Figure 12.** Boxplots of $Q_{pred}/Q_{obs}$ ratios for all clusters and AEPs: (**a**) cluster 1 (C1); (**b**) cluster 2 (C2); (**c**) cluster 3 (C3); (**d**) cluster 4 (C4); and (**e**) cluster 5 (C5).

It is seen in Figures 11 and 12 that cluster 5 is the worst performing group. It is clear that most of the predictions are underestimations. The boxplots are not visible for all the quantiles as there are cases with a gross underestimation. The median of 50% AEP flood is far below the expected line in case of both Figures 11 and 12 for cluster 5. Cluster 5 is made of stations having larger area than the other clusters. Poor performance of cluster 5 in QRT indicates that pooling of larger catchments into single group do not represent a viable choice in RFFA. Cluster 2 seems to be the best performing out of all the five clusters. Cluster 4 showing the best $H_1$ value does not perform as good as cluster 2 in the leave-one-out validation.

Figure 13 shows plots of predicted vs observed flood quantiles for all the AEP floods for the five clusters. These plots generally present a good agreement between the predicted and observed flood quantiles. For cluster 1, there are a few cases of over-estimation when the observed flows are in between 100 $m^3/s$ to 200 $m^3/s$ and in case of larger observations ranging from 1200 $m^3/s$ to 1800 $m^3/s$, there are some under-estimations by the regression model. For smaller discharges, cluster 2 seems to be performing well, although as the discharge gets larger the prediction by the model gets more erroneous, which is also visible from the boxplots. In case of clusters 3 and 4, the models perform well for smaller discharges; for the larger discharges, the models provide gross under-estimation. Cluster 5 is the worst performing group as seen from Figure 13. Cluster 2 performs the best in case of 5% AEP flood. As 5% AEP is the most frequently adopted flood quantile in design flood estimation, it can be said that regions formed based on small to medium sized area with a small range in other catchment characteristics will generate better prediction than other groups. Looking into the homogeneity analysis for all the five clusters, it can be concluded that homogeneity does not play a vital role in enhancing the prediction accuracy.

Tables 9 and 10 show the comparison of mean, median and standard deviation (Std Dev) of absolute relative error ($RE_{abs}$) and absolute $Q_{pred}/Q_{obs}$ ratio for all the clusters and the selected AEPs. Tables 10 and 11 again prove the worst performance by cluster 5 with very large values for both $RE_{abs}$ and $Q_{pred}/Q_{obs}$ ratio values. Clusters 3 and 4 also show a mixture of under- and over-estimation. Clusters 1 and 2 seem to show promising results, although in the case of cluster 1, the mean $RE_{abs}$ and mean $Q_{pred}/Q_{obs}$ ratio values for 1% AEP flood are quite high. Cluster 2 shows the lowest values for both the overall REabs and $Q_{pred}/Q_{obs}$ further confirming the better performance of cluster 2.

**Table 10.** Comparison of mean, median and standard deviation of $RE_{abs}$ for all five clusters (blue indicates the lowest value).

| RE | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|---|
| 50% | Mean_abs | 76.25 | 56.02 | 104.51 | 64.69 | 471.67 |
| | Median_abs | 48.41 | 36.56 | 54.86 | 33.76 | 183.76 |
| | Std Dev_abs | 89.91 | 70.76 | 108.34 | 76.86 | 580.00 |
| 20% | Mean_abs | 89.88 | 36.57 | 71.11 | 82.74 | 542.75 |
| | Median_abs | 43.06 | 25.92 | 42.39 | 39.87 | 424.41 |
| | Std Dev_abs | 210.38 | 29.72 | 89.01 | 113.28 | 317.63 |
| 10% | Mean_abs | 173.56 | 27.91 | 66.26 | 90.36 | 503.40 |
| | Median_abs | 31.84 | 24.90 | 35.48 | 45.83 | 422.76 |
| | Std Dev_abs | 558.54 | 25.44 | 79.88 | 130.68 | 254.32 |
| 5% | Mean_abs | 412.96 | 27.82 | 66.62 | 94.14 | 453.41 |
| | Median_abs | 28.01 | 22.86 | 32.80 | 46.33 | 403.45 |
| | Std Dev_abs | 1500.24 | 24.66 | 76.09 | 140.28 | 208.10 |
| 2% | Mean_abs | 1427.02 | 34.25 | 71.15 | 96.05 | 403.47 |
| | Median_abs | 21.34 | 26.55 | 46.46 | 47.05 | 380.24 |
| | Std Dev_abs | 5427.65 | 30.44 | 82.16 | 145.37 | 170.96 |
| 1% | Mean_abs | 3637.99 | 44.37 | 78.85 | 96.41 | 376.95 |
| | Median_abs | 21.17 | 30.53 | 42.31 | 40.13 | 342.35 |
| | Std Dev_abs | 13984.54 | 35.89 | 94.50 | 146.08 | 154.74 |
| Overall mean | | 969.61 | 37.82 | 76.42 | 87.40 | 458.61 |
| Overall median | | 33.56 | 26.16 | 42.92 | 43.07 | 387.51 |
| Overall Std Dev | | 6121.12 | 39.71 | 87.63 | 125.93 | 313.49 |

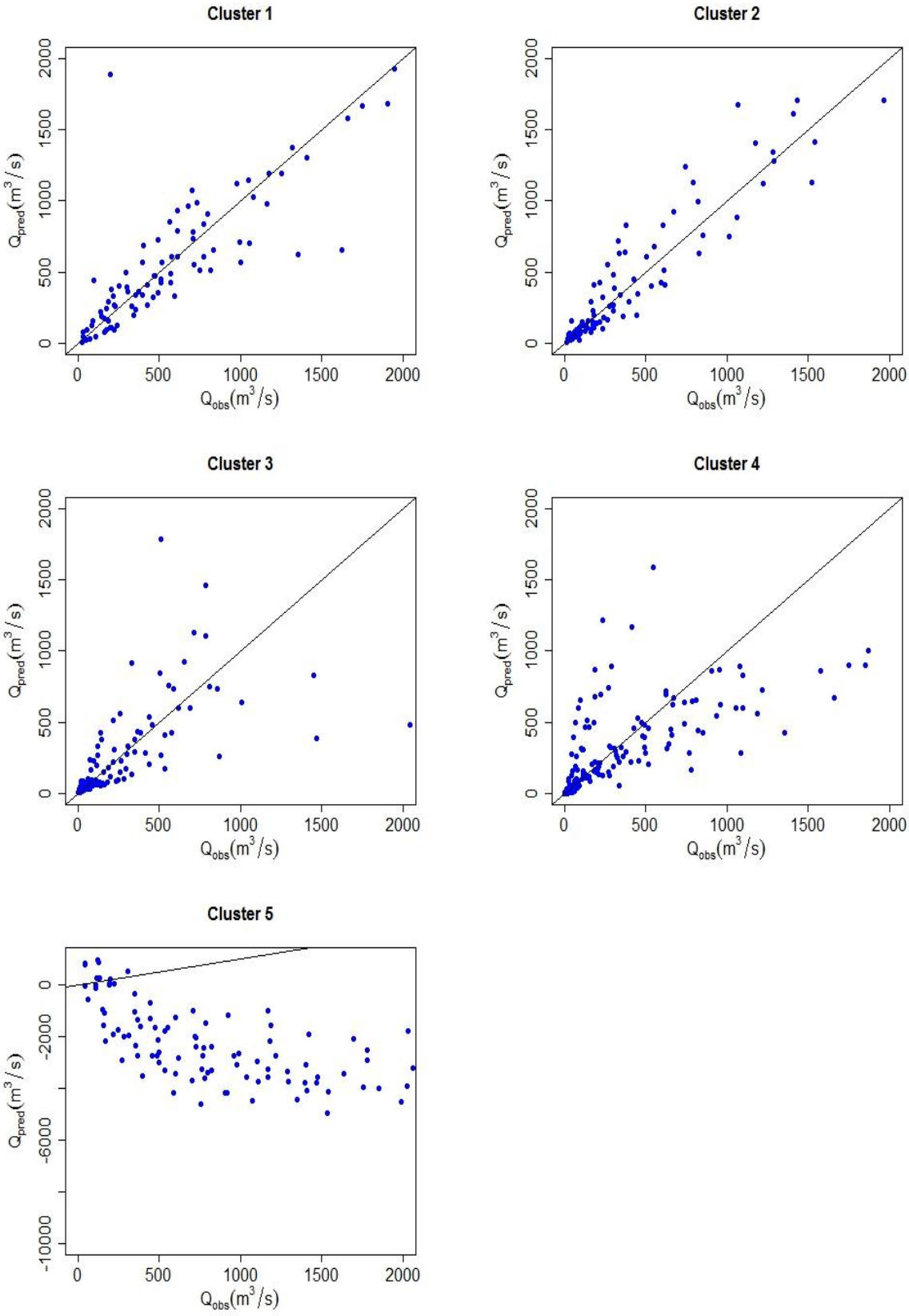


**Figure 13.** Comparison of observed and predicted flood quantiles for all the quantiles and clusters.

**Table 11.** Comparison of mean, median and standard deviation of $Q_{pred}/Q_{obs}$ ratios for all five clusters (blue indicates the lowest value).

| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|---|
| 50% | Mean_abs | 1.38 | 1.21 | 1.52 | 1.35 | 4.68 |
| | Median_abs | 1.08 | 0.97 | 0.78 | 1.00 | 1.85 |
| | Std Dev_abs | 1.13 | 0.89 | 1.43 | 0.95 | 6.10 |
| 20% | Mean_abs | 1.60 | 1.09 | 1.33 | 1.44 | 4.43 |
| | Median_abs | 1.03 | 0.85 | 0.89 | 0.88 | 3.24 |
| | Std Dev_abs | 2.21 | 0.47 | 1.11 | 1.34 | 3.18 |
| 10% | Mean_abs | 2.47 | 1.07 | 1.30 | 1.49 | 4.03 |
| | Median_abs | 1.02 | 1.02 | 0.91 | 0.90 | 3.23 |
| | Std Dev_abs | 5.66 | 0.38 | 1.00 | 1.52 | 2.54 |
| 5% | Mean_abs | 4.88 | 1.07 | 1.31 | 1.52 | 3.53 |
| | Median_abs | 1.04 | 1.01 | 0.98 | 0.89 | 3.03 |
| | Std Dev_abs | 15.07 | 0.37 | 0.98 | 1.62 | 2.08 |
| 2% | Mean_abs | 15.02 | 1.10 | 1.36 | 1.54 | 3.03 |
| | Median_abs | 0.95 | 0.98 | 1.01 | 0.87 | 2.80 |
| | Std Dev_abs | 54.35 | 0.45 | 1.04 | 1.67 | 1.71 |
| 1% | Mean_abs | 37.11 | 1.15 | 1.42 | 1.54 | 2.77 |
| | Median_abs | 0.99 | 0.93 | 1.08 | 0.91 | 2.42 |
| | Std Dev_abs | 139.92 | 0.56 | 1.17 | 1.67 | 1.55 |
| Overall mean | | 10.41 | 1.12 | 1.37 | 1.48 | 3.75 |
| Overall median | | 1.00 | 0.96 | 0.96 | 0.94 | 2.88 |
| Overall Std Dev | | 61.26 | 0.54 | 1.10 | 1.46 | 3.25 |

Table 12 summarizes the evaluation statistics from application of leave-one-out validation with respect to *MSE*, *RMSE*, *BIAS*, *RBIAS*, *RRMSE*, *RMSNE*, $RE_r$, and med_$Q_{pred}/Q_{obs}$ based on observed and predicted flood values for all the five clusters in case of 5% AEP flood. A value close to zero is preferable for *MSE* as zero indicates no error in prediction. However, from Table 12 it is seen that all the *MSE* values for the five clusters are very large, in the range of 25,000 to 35,000,000. The smallest *MSE* is found in case of cluster 2 and the value is 25,309. The range of *RMSE* for all the clusters fall between 159 $m^3/s$ to 5800 $m^3/s$. Cluster 2 shows the lowest *RMSE* with a value of 159 $m^3/s$ proving cluster 2 being the best performing group. Cluster 2 also shows the smallest values in case of *RRMSE*, *RMSNE*, *BIAS*, and *RBIAS*. Cluster 1 has large value for both *BIAS* and *RBIAS* (1480.61 and 388.4, respectively). Clusters 4 and 5 show a large negative *BIAS* and cluster 5 shows a very large negative *RBIAS*. The results here are notably higher than those reported in Durocher et al. [67], Chokmani and Ouarda [68], Chebana et al. [47], and Shu and Ouarda [69]. In Rahman et al. [6], independent component regression was adopted to develop flood prediction equations using the same data set as of this study, where error values are similar to this study. It should be noted that the values of *MSE*, *RMSE*, and *BIAS* depend on catchment size, a larger catchment generally has a larger discharge which is likely to result in higher *MSE*, *RMSE*, and *BIAS*.

The $RE_r$ and med_$Q_{pred}/Q_{obs}$ both show, cluster 2 has the smallest values. Cluster 5 is the worst performing group with a high $RE_r$ and median_$Q_{pred}/Q_{obs}$ (403.44 and −3.03, respectively). Hence, it can be said that group of stations having smaller catchment areas and lower range of other catchment characteristics such as cluster 2 is likely to generate more accurate flood prediction in QRT in the study region.

**Table 12.** Statistical evaluation for cluster 1 for 5% AEP flood (blue indicates best result).

| | MSE | RMSE | BIAS | RBIAS | RRMSE | RMSNE | $RE_r$ | med_$Q_{pred}/Q_{obs}$ |
|---|---|---|---|---|---|---|---|---|
| Cluster 1 | 34034166.24 | 5833.88 | 1480.61 | 388.4 | 2.34 | 15.07 | 28.01 | 1.04 |
| Cluster 2 | 25309.52 | 159.09 | 5.54 | 7.30 | 0.01 | 0.37 | 22.86 | 1.01 |
| Cluster 3 | 58766.74 | 242.42 | 11.39 | 30.91 | 0.04 | 0.99 | 32.79 | 0.98 |
| Cluster 4 | 91284.23 | 302.13 | −47.02 | 51.92 | 0.12 | 1.66 | 45.54 | 0.89 |
| Cluster 5 | 17065693.55 | 4131.06 | −4056.77 | −453.41 | 3.61 | 4.96 | 403.44 | −3.03 |

### *4.4. Comparison with ARR RFFA Model*

An assessment of $RE_r$ values between ARR RFFA model [32] and PCR KNN15, PCR KNN25, and QRT models for cluster 2 is presented in this section. ARR RFFA model is developed using a Bayesian generalized least square based parameter regression technique to estimate regional flood quantiles using Australian flood data [32]. The $RE_r$ values for ARR RFFA model and PCR KNN15, PCR KNN25 and QRT models for cluster 2 are compared in Table 13. It is apparent from Table 13 that, the $RE_r$ values for ARR RFFA model (ranging from 56% to 64%) is greater than PCR KNN15, PCR KNN25 and QRT models for cluster 2 ($RE_r$ values ranging from 42% to 69%, 42% to 60% and 22% to 37%, respectively). ARR RFFA model is developed with data from 558 stations from NSW, Victoria and Queensland [58] and PCR KNN15, PCR KNN25 and QRT models for cluster 2 are developed for 15, 25 and 16 stations from NSW. This may be a possible reason for these differences in *RE* values. However, it is promising to see that the $RE_r$ values from the PCR KNN15, PCR KNN25, and QRT models for cluster 2 are analogous to the $RE_r$ values of ARR RFFA model. From this study, it may be argued that PCR may not be a good choice in RFFA in case of NSW. Moreover, a group of stations with smaller catchment areas such as cluster 2 may generate a better RFFA grouping. Further research with additional catchment characteristics data may enhance the reliability of PCR and cluster analysis based RFFA models in the study region.

**Table 13.** Comparison of absolute *RE* (%) values between ARR RFFA model and PCR KNN15, PCR KNN25, and QRT models for cluster 2.

| AEPs | ARR RFFA Model Absolute RE (%) | PCR_KNN15 Absolute RE (%) | PCR_KNN25 Absolute RE (%) | Cluster 2 Absolute RE (%) |
|---|---|---|---|---|
| 50% | 63.07 | 42.7 | 42.70 | 36.56 |
| 20% | 57.25 | 51.25 | 43.64 | 25.92 |
| 10% | 57.48 | 55.39 | 48.77 | 24.9 |
| 5% | 58.85 | 53.41 | 52.62 | 22.86 |
| 2% | 60.39 | 58.87 | 58.25 | 26.55 |
| 1% | 64.06 | 68.67 | 59.66 | 30.53 |

## 5. Conclusions

A total of 88 stations form NSW, Australia and eight catchment characteristics variables are used in this study to compare regression-based RFFA models. Principal components are derived by applying the principal component analysis on the catchment characteristics data set and a multiple linear regression technique is applied to predict flood quantiles. The first five principal components are selected to be the predictor variables in the regression equations. According to the $R^2$ and $adj\_R^2$ values of the developed regression equations, it is found that the principal component regression based RFFA models perform quite poorly.

The application of cluster analysis resulted into five clusters from the selected 88 stations. Cluster 1 has the smallest catchment areas and larger rainfall intensity, mean annual rainfall, mean annual evapo-transpiration, shape factor, forest, and stream density values. Stations in cluster 2 have smaller sized catchments along with moderate values for rainfall intensity, shape factor, stream density, and mean annual rainfall, although the mean annual evapo-transpiration is the highest in case of

cluster 2. Cluster 3 seems to have all the catchment characteristics uniformly distributed. Cluster 4 is influenced by medium sized catchment area, smaller rainfall intensity, mean annual rainfall, mean annual evapo-transpiration, shape factor, and forest and stream density. The largest values for rainfall intensity, mean annual rainfall, mean annual evapo-transpiration, shape factor, and forest and stream density are found in case of cluster 5, although the other characteristics are quite small with large variance in forest cover. A quantile regression technique is applied to all five clusters with leave-one-out validation. Based on the findings of this study, it can be said that cluster 2 is the best performing group among the selected five clusters. It is also found that a relatively smaller catchment areas and small range of other catchment characteristics in a group is likely to result in more accurate RFFA models in the study region.

It is also found that cluster analysis does not generate any homogeneous groups of catchments for the selected dataset in NSW. Furthermore, the degree of heterogeneity does not have any link with RFFA model performance for the dataset used in this study. The best cluster-based quantile regression model in RFFA is found to be more accurate than the currently recommended ARR RFFA Model in Australia.

**Author Contributions:** Conceptualization, A.S.R. and A.R.; methodology, A.S.R.; software, A.S.R.; validation, A.S.R. and A.R.; formal analysis, A.S.R.; investigation, A.S.R.; data curation, A.S.R.; writing—original draft preparation, A.S.R.; writing—review and editing, A.R.; visualization, A.S.R. and A.R.; supervision, A.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Blöschl, G.; Sivapalan, M.; Wagener, T.; Savenije, H.; Viglione, A. (Eds.) *Runoff prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*; Cambridge University Press: Cambridge, UK, 2013.
2. Ouarda, T.B.M.J.; St-Hilaire, A.; Bobée, B. A review of recent developments in regional frequency analysis of hydrological extremes. *Revue des Sciences de l'eau* **2008**, *21*, 219–232. [CrossRef]
3. Ouarda, T.B.M.J.; Bâ, K.M.; Diaz-Delgado, C.; Carsteanu, A.; Chokmani, K.; Gingras, H.; Quentin, E.; Trujillo, E.; Bobée, A.B. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. *J. Hydrol.* **2008**, *348*, 40–58. [CrossRef]
4. Haddad, K.; Rahman, A.; Ling, F. Regional flood frequency analysis method for Tasmania, Australia: A case study on the comparison of fixed region and region-of-influence approaches. *Hydrol. Sci. J.* **2015**, *60*, 2086–2101. [CrossRef]
5. Ouarda, T.B.M.J. Regional hydrological frequency analysis. In *Encyclopedia of Environmetrics*; El-Shaarawi, A.H., Piegorsch, W.W., Eds.; Wiley: New York, NY, USA, 2013.
6. Rahman, A.S.; Khan, Z.; Rahman, A. Application of Independent Component Analysis in Regional Flood Frequency Analysis: Comparison between Quantile Regression and Parameter Regression Techniques. *J. Hydrol.* **2019**, *581*, 124372. [CrossRef]
7. Acreman, M.C. *Regional Flood Frequency Analysis in the UK: Recent Research-New Ideas*; Institute of Hydrology: Wallingford, UK, 1987.
8. Acreman, M.C.; Sinclair, C.D. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. *J. Hydrol.* **1986**, *84*, 365–380. [CrossRef]
9. Eng, K.; Tasker, G.D.; Milly, P.C.D. An analysis of region-of-influence methods for flood regionalisation in the-Gulf-Atlantic rolling plains. *J. Am. Water Resour. Assoc.* **2005**, *41*, 135–143. [CrossRef]
10. Pilgrim, D.H. *Australian Rainfall and Runoff*; Institution of Engineers: Barton, Australia, 1987.
11. Tasker, G.D.; Hodge, S.A.; Barks, C.S. Region of influence regression for estimating the 50 year flood at ungauged sites. *J. Am. Water Resour. Assoc.* **1996**, *32*, 163–170. [CrossRef]
12. Burn, D.H. An appraisal of the "region of influence" approach to flood frequency analysis. *Hydrol. Sci. J.* **1990**, *35*, 149–165. [CrossRef]
13. Burn, D.H. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resour. Res.* **1990**, *26*, 2257–2265. [CrossRef]

14. Chebana, F.; Ouarda, T.B.M.J. Depth and homogeneity in regional flood frequency analysis. *Water Resour. Res.* **2008**, *44*, W11422. [CrossRef]
15. Hosking, J.R.M.; Wallis, J.R. Some statistics useful in regional frequency analysis. *Water Resour. Res.* **1993**, *29*, 271–281. [CrossRef]
16. Merz, R.; Blöschl, G. Flood frequency regionalisation—Spatial proximity vs. catchment attributes. *J. Hydrol.* **2005**, *302*, 283–306. [CrossRef]
17. Burn, D.H. Cluster analysis as applied to regional flood frequency. *J. Water Res. Plan. Man.* **1989**, *115*, 567–582. [CrossRef]
18. Burn, D.H.; Boorman, D.B. Estimation of hydrological parameters at ungauged catchments. *J. Hydrol.* **1993**, *143*, 429–454. [CrossRef]
19. Himeidan, Y.E.S.; Hamid, E.E.H. Rainfall variability in New Halfa agricultural scheme (Sudan). *Univ. Khartoum J. Agric. Sci.* **2019**, *14*, 383–391.
20. Hughes, J.M.R.; James, B. A hydrological regionalization of streams in Victoria, Australia, with implications for stream ecology. *Mar. Freshw. Res.* **1989**, *40*, 303–326. [CrossRef]
21. Mosley, M.P. Delimitation of New Zealand hydrologic regions. *J. Hydrol.* **1981**, *49*, 173–192. [CrossRef]
22. Nathan, R.J.; McMahon, T.A. Identification of homogeneous regions for the purposes of regionalisation. *J. Hydrol.* **1990**, *121*, 217–238. [CrossRef]
23. Rasheed, A.; Egodawatta, P.; Goonetilleke, A.; McGree, J. A Novel Approach for Delineation of Homogeneous Rainfall Regions for Water Sensitive Urban Design—A Case Study in Southeast Queensland. *Water* **2019**, *11*, 570. [CrossRef]
24. Santos, C.A.G.; Moura, R.; da Silva, R.M.; Costa, S.G.F. Cluster Analysis Applied to Spatiotemporal Variability of Monthly Precipitation over Paraíba State Using Tropical Rainfall Measuring Mission (TRMM) Data. *Remote Sens.* **2019**, *11*, 637. [CrossRef]
25. Tasker, G.D. Comparing methods of hydrologic regionalisation. *J. Am. Water Resour. Assoc.* **1982**, *18*, 965–970. [CrossRef]
26. Hosking, J.R.M.; Wallis, J.R. *Regional Frequency Analysis: An Approach based on L-moments*; Cambridge University Press: New York, NY, USA, 1997.
27. Eng, K.; Milly, P.C.; Tasker, G.D. Flood regionalisation: A hybrid geographic and predictor-variable region-of-influence regression method. *J. Hydrol. Eng.* **2007**, *12*, 585–591. [CrossRef]
28. Eng, K.; Stedinger, J.R.; Gruber, A.M. Regionalisation of streamflow characteristics for the Gulf-Atlantic rolling plains using leverage-guided region-of-influence regression. In Proceedings of the World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat, Tampa, Florida, 15–19 May 2007; pp. 1–11.
29. Gaál, L.; Kyselý, J.; Szolgay, J. Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia. *Hydrol. Earth Sys. Sci. Discuss.* **2008**, *12*, 825–839. [CrossRef]
30. Haddad, K.; Rahman, A. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework: Quantile regression vs. parameter regression technique. *J. Hydrol.* **2012**, *430–431*, 142–161. [CrossRef]
31. Micevski, T.; Hackelbusch, A.; Haddad, K.; Kuczera, G.; Rahman, A. Regionalisation of the parameters of the log-Pearson 3 distribution: A case study for New South Wales, Australia. *Hydrol. Process.* **2015**, *29*, 250–260. [CrossRef]
32. Rahman, A.; Haddad, K.; Kuczera, G.; Weinmann, P.E. Regional flood methods. *Aust. Rainfall Runoff.* **2019**, *3*, 105–146.
33. Zrinji, Z.; Burn, D.H. Regional flood frequency with hierarchical region of influence. *J. Water Res. Plan. Man.* **1996**, *122*, 245–252. [CrossRef]
34. Burn, D.H.; Goel, N.K. The formation of groups for regional flood frequency analysis. *Hydrol. Sci. J.* **2000**, *45*, 97–112. [CrossRef]
35. Castellarin, A.; Burn, D.H.; Brath, A. Assessing the effectiveness of hydrological similarity measures for regional flood frequency analysis. *J. Hydrol.* **2001**, *241*, 270–285. [CrossRef]
36. Burn, D.H. Catchment similarity for regional flood frequency analysis using seasonality measures. *J. Hydrol.* **1997**, *202*, 212–230. [CrossRef]
37. Lim, Y.H.; Lye, L.M. Regional flood estimation for ungauged basins in Sarawak, Malaysia. *Hydrol. Sci. J.* **2003**, *48*, 79–94. [CrossRef]

38. Zrinji, Z.; Burn, D.H. Flood frequency analysis for ungauged sites using a region of influence approach. *J. Hydrol.* **1994**, *153*, 1–21. [CrossRef]
39. Bates, B.C.; Rahman, A.; Mein, R.G.; Weinmann, P.E. Climatic and physical factors that influence the homogeneity of regional floods in south-eastern Australia. *Water Resour. Res.* **1998**, *34*, 3369–3382. [CrossRef]
40. Fill, H.D.; Stedinger, J.R. Using regional regression within IF procedures and an empirical Bayesian estimator. *J. Hydrol.* **1998**, *210*, 128–145. [CrossRef]
41. Haddad, K.; Rahman, A.; Stedinger, J.R. Regional flood frequency analysis using Bayesian generalized least squares: A comparison between quantile and parameter regression techniques. *Hydrol. Process.* **2012**, *26*, 1008–1021. [CrossRef]
42. Griffis, V.W.; Stedinger, J.R. The use of GLS regression in regional hydrologic analyses. *J. Hydrol.* **2007**, *344*, 82–95. [CrossRef]
43. Micevski, T.; Kuczera, G. Combining site and regional flood information using a Bayesian Monte Carlo approach. *Water Resour. Res.* **2009**, *45*. [CrossRef]
44. Ouali, D.; Chebana, F.; Ouarda, T.B.M.J. Quantile regression in regional frequency analysis: A better exploitation of the available information. *J. Hydrometeorol.* **2016**, *17*, 1869–1883. [CrossRef]
45. Rahman, A.; Charron, C.; Ouarda, T.B.M.J.; Chebana, F. Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Stoch. Environ. Res. Risk A* **2018**, *32*, 123–139. [CrossRef]
46. Rahman, A. A quantile regression technique to estimate design floods for ungauged catchments in south-east Australia. *Australas. J. Water Resour.* **2005**, *9*, 81–89. [CrossRef]
47. Chebana, F.; Charron, C.; Ouarda, T.B.M.J.; Martel, B. Regional frequency analysis at ungauged sites with the generalized additive model. *J. Hydrometeorol.* **2014**, *15*, 2418–2428. [CrossRef]
48. Burn, D.H. Delineation of groups for regional flood frequency analysis. *J. Hydrol.* **1988**, *104*, 345–361. [CrossRef]
49. DeCoursey, D.G.; Deal, R.B. General Aspects of Multivariate Analysis with Applications. *Misc. Publ.* **1974**, *1275*, 47.
50. Hawley, M.E.; McCuen, R.H. Water yield estimation in western United States. *J. Irrig. Drain. Div.* **1982**, *108*, 25–34.
51. Kar, A.K.; Goel, N.K.; Lohani, A.K.; Roy, G.P. Application of clustering techniques using prioritized variables in regional flood frequency analysis—Case study of Mahanadi Basin. *J. Hydrol. Eng.* **2011**, *17*, 213–223. [CrossRef]
52. Choi, T.H.; Kwon, O.E.; Koo, J.Y. Water demand forecasting by characteristics of city using principal component and cluster analyses. *Environ. Eng. Res.* **2010**, *15*, 135–140. [CrossRef]
53. Haque, M.M.; de Souza, A.; Rahman, A. Water demand modelling using independent component regression technique. *Water Resour. Res.* **2017**, *31*, 299–312. [CrossRef]
54. Haque, M.M.; Rahman, A.; Hagare, D.; Kibria, G. Principal component regression analysis in water demand forecasting: An application to the Blue Mountains, NSW, Australia. *J. Hydrol. Environ. Res.* **2013**, *1*, 49–59.
55. Koo, J.Y.; Yu, M.J.; Kim, S.G.; Shim, M.H.; Koizumi, A. Estimating regional water demand in Seoul, South Korea, using principal component and cluster analysis. *Water Sci. Tech. Water Supply* **2005**, *5*, 1–7. [CrossRef]
56. Ball, J.; Babister, M.; Nathan, R.; Weeks, W.; Weinmann, P.E.; Retallick, M.; Testoni, I. *Australian Rainfall and Runoff-A Guide to Flood Estimation*; Engineers Australia: Canberra, Australia, 2019.
57. Rahman, A.; Haddad, K.; Haque, M.; Kuczera, G.; Weinmann, P.E. *Australian Rainfall and Runoff Project 5: Regional Flood Methods: Stage 3 Report*; (No. P5/S3, p. 025). technical report; Engineers Australia: Canberra, Australia, 2015.
58. Çamdevýren, H.; Demýr, N.; Kanik, A.; Keskýn, S. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. *Ecol. Model.* **2005**, *181*, 581–589. [CrossRef]
59. Olsen, R.L.; Chappell, R.W.; Loftis, J.C. Water quality sample collection, data treatment and results presentation for principal components analysis–literature review and Illinois River watershed case study. *Water Res.* **2012**, *46*, 3110–3122. [CrossRef] [PubMed]
60. Pires, J.C.M.; Martins, F.G.; Sousa, S.I.V.; Alvim-Ferraz, M.C.M.; Pereira, M.C. Selection and validation of parameters in multiple linear and principal component regressions. *Environ. Modell. Softw.* **2008**, *23*, 50–55. [CrossRef]

61. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*; PrenticeHall International. Inc.: New Jersey, NJ, USA, 2007.
62. Baeriswyl, P.A.; Rebetez, M. Regionalization of precipitation in Switzerland by means of principal component analysis. *Theor. Appl. Climatol.* **1997**, *58*, 31–41. [CrossRef]
63. Bhaskar, N.R.; O'Connor, C.A. Comparison of method of residuals and cluster analysis for flood regionalization. *J. Water Resour. Plan. Manag.* **1989**, *115*, 793–808. [CrossRef]
64. Dinpashoh, Y.; Fakheri-Fard, A.; Moghaddam, M.; Jahanbakhsh, S.; Mirnia, M. Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. *J. Hydrol.* **2004**, *297*, 109–123. [CrossRef]
65. Rao, A.R.; Srinivas, V.V. Regionalization of watersheds by hybrid-cluster analysis. *J. Hydrol.* **2006**, *318*, 37–56. [CrossRef]
66. Kuczera, G. *FLIKE HELP*; Chapter 2 FLIKE Notes; University of Newcastle: Callaghan, Australia, 1999.
67. Durocher, M.; Burn, D.H.; Zadeh, S.M. A nationwide regional flood frequency analysis at ungauged sites using ROI/GLS with copulas and super regions. *J. Hydrol.* **2018**, *567*, 191–202. [CrossRef]
68. Chokmani, K.; Ouarda, T.B.M.J. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. *Water Resour. Res.* **2004**, *40*. [CrossRef]
69. Shu, C.; Ouarda, T.B.M.J. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resour. Res.* **2007**, *43*. [CrossRef]