An estimation of regional and at-site quantiles of extreme winds under flood index procedure

Ishfaq Ahmad
International Islamic University, Islamabad

Touqeer Ahmad (touqeer.ahmad@studenti.unipd.it)
University of Padova

Ibrahim Mufrah Almanjahie
King Khalid University

Muhammad Athar Ameer
International Islamic University

Research Article

Keywords: Linear-moments, Monte Carlo simulation, Quantile estimates, Wind speed, Regional frequency analysis

Posted Date: June 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1668488/v1

An estimation of regional and at-site quantiles of extreme winds under flood index procedure

Ishfaq Ahmad[1], –Touqeer Ahmad*[2], –Ibrahim Mufrah Almanjahie[3, 4]– Muhammad Athar Ameer[1]

[1] *Department of Mathematics and Statistics, Faculty of Basic and Applied Sciences, International Islamic University, 44000 Islamabad, Pakistan.*

[2] **Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121, Padova, PD, Italy.*

[3] *Department of Mathematics, King Khalid University, Abha 62529, Saudi Arabia.*

[4] *Statistical Research and Study Support Unit, King Khalid University, 62529, Abha, Saudi Arabia.*

***Corresponding author:** Touqeer Ahmad Ph.D. scholar, at the department of statistical science University of Padova, Italy. (email. touqeer.ahmad@studenti.unipd.it)*

Abstract: Daily annual maximum wind speed values are among environmental events with the most catastrophic consequences for society. A comprehensive understanding of the magnitudes and frequency of extreme winds is essential for ensuring sustainably managed wind energy, planning for emergency weather situations, and designing different building structures. This study investigated regional frequency analysis of daily annual maximum winds for Baluchistan and Sindh provinces of Pakistan. We intended to find the regional frequency distribution for maximum winds and predict the returns for extreme winds events in the future. L-moments regionalization techniques along with flood index procedure were applied to maximum wind speed records of 21 stations across the study area. Only one station namely Lasbella was found to be discordant. With the help of cluster analysis, the remaining twenty stations were further divided into two homogeneous. Heterogeneity measures validate that both regions are homogenous with allotted stations. Regional quantiles for both regions are estimated through

best-fit probability distribution among Generalized Normal (GNO), Generalized Logistic (GLO), Pearson Type 3 (P3), Generalized Pareto (GPA), and Generalized Extreme Value (GEV). Robustness of GLO distribution compared to GEV distribution is assessed through Monte Carlo simulations of relative bias and relative root mean square error. Findings clearly show that GLO distribution is best for the regional modeling. Furthermore, at-site quantiles are estimated by multiplying quantities of regional quantiles using the sample mean and median as scaling factors. Estimated quantiles can be helpful for future planning concerning wind energy and the codification of structural designs and policy implications.

Keywords: Linear-moments, Monte Carlo simulation, Quantile estimates, Wind speed, Regional frequency analysis

1 Introduction

In dry and semi-arid regions of the world, wind is considered a prominent environmental phenomenon. Wind storms is a major agricultural, ecological and urban population problem that has gotten worse in recent years. Accurate codification of extreme wind speeds events is required in various design projects such as buildings, bridges, wind turbines and tower installation. Correct estimations are helpful to ensure the safety and reliability of the structures, and to save the agricultural farms and human beings. On the other hand, wind also plays a significant role in the wind power generation field due to its reasonable efficiency and suitability for the environment. Pakistan is located in the monsoon zone and is greatly affected by wind storms. For these reasons, the frequency of extreme winds are sometimes quantified in terms of extreme quantiles or return levels x_T, which is the maximum wind speed that is likely to be exceeded once every T-years on average.

In most cases, the frequency analysis of extreme wind speed through quantiles are intended by fitting theoretical statistical distributions to a sample of observed wind speed data. Frequency analysis of wind offers valuable information for individuals working in the field of structural and environmental design and renewable energy studies (Lun and Lam, 2000). In the perspective of wind speed

modelling, the Weibull distribution has been recommended widely in literature (Nedaei et al. 2014; Tizpar et al. 2014; Weisser 2003; Ohunakin and Akinnawonu 2012).) In addition, the Gumbel and generalized extreme value (GEV) distributions are also commonly used for the analysis of extreme winds (Wang et al. 2015; Lee et al. 2012; Lombardo et al. 2009; Yao et al. 2012). There is no consent on which distribution is suitable, and the performance differs in different areas. Consequently, numerous extreme value models are incessantly debated and enhanced, for example, the multivariate compound extreme value distribution (Liu et al. 2006) peak-over-threshold (Liang et al. 2019).

Pakistan, an emerging state, faces many devastations due to extreme wind speed, which causes financial loss, physical demolitions, and human and animal injuries. Wind speed modelling is essential for the country to overcome destructions linked with the extreme wind events. In this study, regional frequency analysis with at-site frequency analysis quantiles are practiced for the estimation of extreme events. The major drawback of linear moment procedure based at-site frequency analysis (ASFA) is that it suffers from sampling inconsistency and specifically for estimation of quantiles for large return periods (Cunnane,1988; Hosking and Wallis, 1993; Fawad et al. 2018). On the other hand, by using the linear moments method based regional frequency analysis (RFA), one can acquire efficient estimates of extreme quantiles which could be used for policy implications (Coles et al. 2001; Fawad et al. 2018).

Several international studies have been conducted in this direction from the last two decades, see, for instance, (Cook and Harris, 2001; Goel et al. 2004; Harris, 2005; Modarres, 2008; Darbandi et al. 2012; Hong et al. 2013; Hong and Ye, 2014; Ghazijahani and Lee, 2015; and Campos and Soares, 2016). The same procedure had also been practiced by (Hosking and Wallis, 1997; Parida et al. 1998; Fowler and Kilsby, 2003; Yue and Wang 2004; Abida and Ellouze, 2007; Seckin et al. 2011; Ahmad et al. 2016 and Komi et al. 2016). The previous related work in

Pakistan includes (Hussain and Pasha, 2009; Shahzadi, 2013). More relevantly, (Fawad et al. 2018) used seven distributions that analyze annual maximum wind speed and used different standards to determine the most suitable distribution for nine stations of Punjab province, Pakistan. In addition, (Ahmad et al. 2019; Ahmad et al. 2021) used the ASFA procedure with a Bayesian framework for modelling extreme rainfall over the country. No study has been conducted before in the provinces, namely Sindh and Baluchistan, using ASFA and RFA approaches. More formally, the importance of these regions for this study is due to their linkage with coastal line.

The paper is organized as follows. Section 2 deal with the materials and methods. For instance, details regarding study area and observed data with their exploratory analysis are given in this section. Further, the procedure of regionalization is also presented in the same section. In Section 3, the results are discussed briefly. Conclusions and some recommendations are given in the final Section 4.

2 Material and methods

2.1 Study areas information

At the initial, this study used the daily annual maximum wind series (DAMWS) of twenty-one sites of Baluchistan and the Sindh provinces of Pakistan. Freely accessible DAMWS for all stations at height ten meters/s were downloaded from the NASA website. Daily maximum wind speed data consists of measurement from 00h to 24h throughout the day. Figure 1 (left) shows the geographical locations of observed sites. We are interested in forming the homogenous regions by using stations' characteristics (e.g., geographical closeness).

[Fig. 1 Place here]

The descriptive statistics corresponding to all stations are explained in Table 1. Figure 1 (right) shows the distribution (so-called the ridgeline) plots of the

DAMWS for observed stations with respective province in parentheses. It can be observed from Fig. 1 (right) that the wind speed sometimes at some locations crosses the 20 m/s. The maximum wind speed at Turbat and Punjgur stations are recorded more than 15 m/s many time. Also, results of Table 1 indicate that the wind speed on average is higher at Turbat and Punjgur stations. These may happen due to the existence of these stations near the coastal line area of Baluchistan province.

The basic assumptions (e.g., homogeneity, independence and stationarity) of the data were checked by using the Mann-Whitney U test, Wald-Wolfowitz test, and Spearman rank correlation test, respectively. The finding relevant to fundamental assumptions is given in the results and discussion section. Overall, the tests' results demonstrate that the data meet the basic prerequisites and can be used for further analysis.

[Table 1 Place here]

2.2 Linear moments

Initially, Hosking (1990) developed linear moments (LMs) for sample data and distributions through the expectation of a linear combination of order statistics. In a computational point of view, LMs are considered superior to ordinary moments. The shape of the distributions can be Judged through LMs. LMs also offer the measure of dispersion, location, skewness and kurtosis. In general, they work more efficiently when dealing with small sample sizes and outliers than the other methods (Hosking, 1990; Hosking and Wallis, 1997; Alam et al. 2016; Fawad et al. 2018).

Hosking (1990) defined the LMs in term of linear combination of probability weighted moments (PWMs). Let Y be random variable with cumulative distribution function $H(.)$, the PWMs are defined as

$$\Lambda_r = E[Y\{H(Y)\}^r], \qquad r = 0,1,2,... \tag{1}$$

The LMs are defined as

$$\kappa_{r+1} = \sum_{j=0}^{r} Z_{r,j}\Lambda_r, \qquad r = 0,1,2,...,(n-1) \tag{2}$$

where $Z_{r,j} = (-1)^{r-j}\binom{r}{j}\binom{r+j}{j} = \frac{(-1)^{r-j}(r+j)!}{(k!)^2(r-j)!}$. The first four moments in term of PWMS are given as

$$\Lambda_1 = \kappa_0$$

$$\Lambda_2 = 2\kappa_1 - \kappa_0$$

$$\Lambda_3 = 6\kappa_2 - 6\kappa_1 + \kappa_0$$

$$\Lambda_4 = 20\kappa_3 - 30\kappa_2 + 12\kappa_1 - \kappa_0$$

where κ_1 and κ_2 are the measure of location and dispersion, respectively. The linear moments based measures of coefficient of variation, skewness and kurtosis are defined as:

Coefficient of variation: $\Pi = \kappa_2 / \kappa_1$

Skewness: $\Pi_3 = \kappa_3 / \kappa_2$

kurtosis: $\Pi_4 = \kappa_4 / \kappa_2$

In general, we used unbiased estimator λ_r of Λ_r of PWMS, that is

$$\lambda_r = \frac{1}{n}\sum_{k=r+1}^{n} \frac{(k-1)(k-2)...(k-r)}{(n-1)(n-2)...(n-r)} y_{k:n} \tag{3}$$

The first four sample LMs with combination of sample PWMs are given in the following forms:

$$l_1 = \lambda_0$$

$$l_2 = 2\lambda_1 - \lambda_0$$

$$l_3 = 6\lambda_2 - 6\lambda_1 + \lambda_0$$

$$l_4 = 20\lambda_3 - 30\lambda_2 + 12\lambda_1 - \lambda_0$$

The sample based LMs ratios are defined as $\pi = l_2/l_1$, $\pi_3 = l_3/l_2$, $\pi_3 = l_4/l_2$ .Where π, π_3 and π_3 are sample based linear measure of coefficient, skewness and kurtosis, respectively.

2.3 Linear moments based regional frequency analysis

According to (Hosking and Wallis, 1997), the four steps are required for LMs based RFA. The following are steps of RFA:

Step 1. Data cleaning and screening

Step 2. Construction of homogenous region on the basis of stations characteristics or by using statistical test.

Step 3. Selection of best fit probability distribution.

Step 4. Parameters estimation of the best fitted models and accuracy of estimated quantiles.

On the other hand, the ASFA is also used in this study, the step 2 is not required when we are dealing with ASFA.

2.3.1 Data cleaning and screening

Data screening is used as a tool to remove the inconsistent station from the study. The LMs based discordancy measures play a crucial role in the data cleaning and screening. Discordancy measure called (Φ_i) is defined as

$$\Phi_i = \frac{1}{3}(\nu_i - \bar{\nu})^T Q^{-1}(\nu_i - \bar{\nu}), \tag{4}$$

where $\nu_i = [\pi^{(i)} \ \pi_3^{(i)} \ \pi_4^{(i)}]^T$ is a vector of LMs ratios, and Q is variance covariance matrix. For a decision point of view, we shall remove the station from the analysis if the resultant value of (Φ_i) exceeds the critical value. In this study, the critical value for (Φ_i) is set as 3 for all stations. Screening of data is an early try to form homogenous regions.

2.3.2 Formation of homogenous regions

The formation of homogeneous regions is the most crucial step in RFA. When all the stations in a region share some common characteristics such as sewerage zone, mean yearly rainfall, speed of wind, latitude, longitude, drainage area, and time of occurrences of most significant flood in the year etc., then the region is said to be homogeneous. Many studies have been conducted on different grouping methods, such as geographical convenience, subjective partitioning, objective partitioning, and cluster analysis. In this study, the regions have been constructed by using cluster analysis and heterogeneity test. Cluster analysis is a multivariate statistical approach used to construct the homogenous. It works by organizing stations into groups, or clusters, on the basis of how closely related they are. Here, we use Ward's method to perform the cluster analysis, interested readers can see for example, (Randriamihamison et al. 2021). For higher number of stations in each group/cluster and smaller range of drainage area, the cluster analysis procedure is considered more appropriate as affirmed by (Shiau and Wu, 2009; Komi et al.

2016). Second possibility is to compute the heterogeneity measure H of the region which can be constructed on the basis of cluster analysis. Heterogeneity measure H is defined as

$$\mathrm{H} = \frac{\eta - \mu_\eta}{\sigma_\eta}, \qquad (5)$$

Where $\eta = \left[\sum_{i=1}^{N} n_i (\pi^i - \pi^R)^2 \Big/ \sum_{i=1}^{N} n_i \right]^{\frac{1}{2}}$ and $\pi^R = \sum_{i=1}^{N} n_i \pi^i \Big/ \sum_{i=1}^{N} n_i$ is the weighted standard deviation and regional mean of linear moment based sample coefficient of variation. To determine the heterogeneity, the relative dispersion of linear moment-based CVs of each station is measured using weighted standard deviation as shown in the above expression. In addition, a simulation experiment with four-parameter kappa distribution is performed to the heterogeneity of proposed regions. In detail, the kappa distribution is fitted to sample LMs ratios, and the distribution parameters are estimated analytically. Then, the sample of size 500 is a simulation from the fitted distribution (i.e. Kappa distribution) for each region. The experiment is repeated 10^2 times. Mean and variance is calculated corresponding to every stations and is denoted by μ_{η_j} and σ_{η_j}. Now, it's time to look at H. If H< 1 refers to an adequately homogeneous region, if $1 \leq \mathrm{H} < 2$ is possibly homogeneous region and if H >2 the region will become heterogeneous.

2.4 Selection of an appropriate distribution

An appropriate model selection is required for constructed homogenous regions. Hosking and Wallis (1997) argue that best-fit distribution quantile estimates are more accurate and reasonable. Generally, the distribution with the higher number of parameters comprises less bias and provide a suitable estimation of extreme quantiles. Therefore, the goodness of fit measure is defined as:

$$Z^{dist} = \frac{\pi_4^{dist} - \pi_4^R + B_4}{\sigma_4}, \qquad (6)$$

where π_4^{dist} and π_4^R are fitted distribution L-kurtosis and : regional weighted average L-kurtosis, respectively. B_4 and σ_4 are the bias and standard deviation of π_4^R. The smaller value of Z^{dist} is the sign of good fit, which indicates that the true distribution is same as the distribution we are fitting to the data. If the value of Z^{dist} is about close to zero or if $\left|Z^{dist}\right| \leq 1.645$ that is, the fit is recognized to be appropriate

2.5 Inference and robustness of estimated quantiles

This section deals with parameters estimation of the selected appropriate model and evaluation of its accuracy in producing efficient quantiles estimates for all sites in the homogeneous region. (Hosking and Wallis, 1997) suggested that the regional LMs algorithm be more practical despite the non-fulfilment of some basic assumptions of the index flood procedure. Simulation based regional quantiles are estimated for several non-exceedance probabilities. Additionally, the quantile estimates of each site could be obtained by scaling $Q(.)$ with an estimate of the scaling factor of μ_i conforming to non-exceedance probability G as follows:

$$\hat{q}(G) = l_1^{(i)} \hat{Q}(G), \qquad (7)$$

where $Q(.)$ is the quantile function of regional frequency distribution, $l_1^{(i)}$ is the sample mean. The robustness of designated regional frequency distribution is relative bias (RB) and relative root mean square error (RRMSE). We implemented the Monte Carlo simulation technique with 10,000 simulations to estimate RB and RRMSE. For more details about assessment measures, we refer to (Fawad et al. 2018). The details findings are given on the results and discussion section.

3. Results and Discussion

3.1 Basic Assumptions

It is necessary to test the fundamental assumptions of the observed data before conducting RFA. On the other hand, the final results could be doubtful without satisfying the basic assumptions of the data. Wald-Wolfowitz, Mann-Whitney, and Spearman's rank correlation tests are used to test the assumptions of independence, homogeneity, and stationarity. For the theoretical background of these procedures, we refer to (Naghettini, 2017, chap. 7). In Table 2, the $p > 0.05$ corresponding to each station indicate that the basic assumptions of the data are satisfied and the data of the considered stations can be used for RFA.

[Table 2 Place here]

Further to explore the configuration of the DAMWS, time series plots are also constructed. The pattern of DAMWS observation from 1990 to 2019 is the confirmation of accepting the null hypothesis of stationarity. The time series plots corresponding to Karachi and Thatta stations are depicted in Fig. 2. From the plots, it can be noted that there is no upward and downward pattern in the data. Similarly, the other sites also having the similar pattern to these two.

[Fig. 2 place here]

3.2 Screening of the data using Discordancy Measure

In RFA, data screening is considered the essential step for discordant stations if any; for this, the most suitable discordancy measure statistic Φ_i is used. The discordancy statistic Φ_i is estimated for every station by assuming all stations lie in one region. The results are given in Table 3. It can be noted that the calculated values of Φ_i for all sites are less than the critical value (i.e., 3) except Lasbella

which has computed value 3.05. Therefore, the Lasbella station is declared as discordant. Similar statistic measures have been by many researcher (see for instance, Fawad et al., (2018) and references therein).

In addition, some other quantities are also reported in Table 3. For instance, n is the number of observed in respective station. Further, l_1, π, π_3 and π_4 indicate the first sample moment (i.e., sample mean), L-CV, linear skewness and linear kurtosis, respectively. The Turbat station has the highest mean with relatively low L-CV in the data. Most stations have positively skewed data, while three stations, namely Hyderabad, Larkana and Rohri have negatively skewed data.

[Table 3 place here]

3.3 Regions formation

After removing the discordant station from the data, the next step is to construct the homogeneous regions. The term "homogeneous region" implies the sites alliance into homogenous regions by their mutual characteristics, such as geographical, hydrological, or others. In this study, we use heterogeneity measure H and cluster analysis for developing the regions. By following the procedure given in (subsection 2.3.2), H_1, H_2 and H_3 heterogeneity measures based on L-Cv, L-Skewness and L-Kurtosis are calculated by keeping all stations in one region. Table 4 display the results of heterogeneity measures.

[Table 4 place here]

The statistic value of $H_1 < 1$ and $H_3 < 1$ represent that the all station forming exact homogenous region. Further, the statistic value of $H_2 < 2$ also indicate that all the stations might be existing in one homogenous region. The statistic $H_2 > 1$ also suggesting to construct more than one homogenous region. For limits interpretation

of H statistic (see for example subsection 2.3.2). For further clarification about the formation of homogenous region we perform the hierarchical cluster analysis.

[Fig 3 place here]

Hierarchical clustering is considered a very efficient tool for developing homogenous region. Figure 3 displays the dendrogram constructed trough hierarchical cluster analysis (Ward's method) using the information of all 20 stations. Only two clusters are identified in Fig. 3, cluster one (called Region I) have 13 stations and cluster two (called Region II) have 7 stations. The LMs based results of the stations inside the both regions are also provided Table 3. Further, Region I shows a relatively higher slope. To verify the homogeneity measure inside the cluster and/or the regions, again LMs ratios and discordancy measures Φ_i are calculated separately for each station in both regions. Table 5 reports the results of stations involved in both regions. The results clearly show that no station is discordant in both regions. This means that these stations can be used for RFA and quantile estimation. Furthermore, the heterogeneity measures H_1, H_2 and H_3 are again estimated for both regions individually. Results are shown in Table 5. Table 6 validates our constructed regions and shows that both regions are homogenous. After developing the homogenous, the next step is to decide the best-fitted model for each region. The details regarding the goodness of fit of the probability model are given in a subsequent section.

[Table 5 place here]

[Table 6 place here]

3.5 Selection of Best-Fitted Regional Frequency Distribution

Selection of best fit distribution in RFA has always been of primary concern to meteorologists. Besides finding the best fit distribution, we also aim to provide distributions with precise quantiles' estimates for different return periods. Initially, graphical evaluation and a goodness of fit measure (i.e., ZDIST statistic) suggested

by Hosking and Wallis (1997) are used to select the best fitted distribution among five distributions namely Generalized Logistic (GLO), Generalized Pearson type III (PE3), Generalized Extreme Value (GEV), Generalized Normal (GNO), and Generalized Pareto (GPA). According to Hosking and Wallis (1997, If the critical value i.e., $|Z^{Dist}| \leq |Z_{0.05}| = 1.64$, we can reject the hypothesis of homogeneity. The results linked with Z^{Dist} statistic are given in Table 7. Hence, the Z^{Dist} statistic indicate that the GLO distribution is an appropriate for modelling wind data in both regions. Furthermore, the second best distribution to region one is GEV and to region two is GNO. For more clarification, we perceive fitting through LMs ratio diagram.

[Table 7 place here]

LMs ratio diagram is a graphical tool which is very popular to select best model extreme value theory. In the LMs ratio diagram (see Fig. 4), the red dot sign represents the sample regional skewness and kurtosis ratio as shown on the curve of the distributions. The red dot is existing very close to GLO distribution curve in both regions. This clearly indicate that the GLO distribution is more appropriate for DAMWS in both regions.

[Fig. 4 place here]

3.7 Parameter estimates and regional quantiles

This section deals with the estimation of regional parameters of the fitted distribution. Also, the regional quantiles corresponding to different return periods (T) are estimated by using regional parameters estimates of the distribution in the quantile function of the distribution. Further, the return periods can be written in terms of exceedances probability (i.e., p=1/T). On the other hand, the non-exceedances probability regarding period T is p=1-1/T. The results related to parameter estimates and return levels for both regions are shown in Table 8.

[Table 8 place here]

Additionally, the regional growth curve with error bounds for GLO distribution has been produced for both regions. The regional growth curve for any distributions is pretentious by the magnitude of L-CV and L-skewness Fawad et al. (2018). It is shown in Figs. 5 (a, b) that GLO distributions have a good slope. It is portrayed in Fig. 5 that the growth curves with error bounds of GLO for both regions have similar behaviour up to 20 years return periods. Errors bounds for the GLO growth curve in both regions at a higher level are relatively narrow.

[Fig. 5 place here]

To choose most reliable distribution among the distributions standing at first and second position in Table 7. We practiced a Monte Carlo simulation algorithm proposed by Meshgi and Khalili (2009) for design flood estimates to assess the efficiency of RFA distribution through relative bias (RB) and relative root mean square error (RRMSE). The simulation results of RB and RRMSE corresponding to different return periods are shown Fig. 6. Figures 6 (a, b) is representing the region I and it show that GLO distribution have smaller RB and RRMSE for the all return periods than GEV distribution. On the other hand, Figs. 6 (c, d) represent RB and RRMSE related to region II. Similar to region I, the GLO distribution again having the lowest RB and RRSME than GNO distribution. Overall, Fig. 6 specify that that GLO distribution is outperforming in both regions than GNO and GEV distribution.

[Fig. 6 place here]

3.8 At-site quantiles estimation

It is not possible in practice to use regional quantiles until we calculate at-site quantiles. For each station, at-site quantiles can be derived by multiplying the regional quantiles by the sample mean or median of that particular station, i.e, $\hat{q}(G) = \ell_1^{(i)} \hat{Q}_i(G)$, where G=1/T.

[Table 9 place here]

In this regional study, estimates of at-site quantiles are produced for both regions using the quantile function of best-fitted distribution to regions (i.e., GLO) while sample means and medians playing a role as scale factors. The result of estimated quantile for region I and region II are presented in Table 9 and Table 10. From the findings given in Table 8 and Table 9, we notice that the quantiles estimate corresponding to a region I are slightly higher for some stations, namely Mirpurkhas, Thatta and Karachi airport when using the median as a scaling factor. Similarly, the quantile estimates related to the Rohri station are slightly higher for smaller return periods when the median plays a scaling factor. This may happen due to higher median than mean. On the other hand, all the stations of both regions have higher quantiles when using mean as scale factor except Mirpurkhas, Thatta, Karachi airport and Rohri (see Table 9 and 10).

[Table 10 place here]

In addition, the standard errors of at-site quantiles estimate for both regions are calculated by using the following equation.

$$Var(\hat{Q}(G) \approx \{x(G;\theta_0)\}^2 Var(\hat{\mu}_i)) + \mu_i^2 Var\{x(Q;\theta_0)\} \qquad (8)$$

In practice, the term $x(G;\theta_0)$ is substituted by estimates of regional quantiles μ_i, with a sample scaling factor, which might be sample mean or sample median Fawad et al (2018). The results of standard errors for station of both regions are shown in Tables 11 and Table 12 using mean and

median as scaling factors, respectively. To determine which scaling factor is relatively superior, we can compare Tables 11 and 12 together. Overall, the standard errors for region I and region II based on mean scaling factor are relatively smaller than the standard errors based on median scaling factor.

[Table 11 place here]

[Table 12 place here]

4 Conclusions and Recommendations

This study investigates the RFA of DAMWS observed at twenty-one sites over the Baluchistan and Sindh provinces of Pakistan. The results confirm that the wind records fulfill the underlying assumptions of independence, homogeneity, and stationarity. Discordancy measure recommends that no other station is discordant except Lasbella station. Therefore, except Lasbella station, all other stations can be considered for further regional frequency analysis. The initial analysis considers the twenty stations to be in one homogeneous region based on their geographical locations. Regional heterogeneity test further demonstrates its homogeneity, which shows that the nine stations constitute a single homogenous region. In addition, the cluster analysis suggests that all twenty station formed two homogenous regions. The best fit distribution has been decided for both regions via ZDIST statistic criteria and L-moments ratio diagram. We observe that the GLO distribution is best fitted to the both homogeneous regions out of the several probability distributions. The robustness of GLO distribution as compared with GEV distribution is also assessed by using RB and RRMSE assessments measures. The results imply that RB and RRMSE are smaller for GLO distribution. Hence, GLO distribution is considered an excellent choice for regional analysis of wind records.

As a second step, we calculated the at-site quantiles by multiplying the regional quantiles with sample means and medians as scaling factors. This procedure called index flood procedure. For at-site quantiles, the performance of GLO distribution is quite reasonable with both scaling factors. In addition, we calculated the standard errors of these at-site quantiles. The findings can be compared with the quantiles and respective standard errors calculated by through at-site frequency analysis for both regions. We notice that the quantiles estimate corresponding to a region I are slightly higher for some stations, namely

Mirpurkhas, Thatta and Karachi airport when using the median as a scaling factor. Similarly, the quantile estimates related to the Rohri station are slightly higher for smaller return periods when the median plays a role as scaling factor in region II. Overall, the standard errors for region I and region II based on mean scaling factor are relatively smaller than the standard errors based on median scaling factor. In order to diminish losses because of heavy wind speeds, the quantiles of DAMWS estimated through quantile function of GLO distribution can be used for policy implications in codifying the wind load for different standardized structural designs. Also, the finding this study could be helpful for engineers in the installation of wind turbine in these considered areas or at ungagged sites.

Acknowledgments: The authors would like to express their gratitude to Department of Statistical Sciences, University of Padova, Italy for giving administrative and technical support. Further, the authors are also grateful to International Islamic University, Islamabad, Pakistan and King Khalid University, Abha, Kingdom of Saudi Arabia.

Author's contribution: Ishfaq Ahmad and Touqeer Ahmad analyzed the raw data and revised the original draft; Ibrahim Mufrah Almanjahie and Muhammad Athar Ameer performed the data analysis and created the figures and prepared the original draft.

Funding: The University of Padua supports this research through library systems.

Code availability: R codes used in this study can be acquired from the corresponding author on reasonable request.

Data availability: The data used in this study is freely available and can be downloaded from NASA website. https://power.larc.nasa.gov/data-access-viewer/.

Declarations

Ethics approval: There is no ethical conflict by all authors.

Consent to participate: All authors agree to participate.

Consent for publication: All authors agree with the publication.

Conflict of interest: The authors declare no competing interests.

REFERENCES

Abida, H. and Ellouze, M. (2007). Probability distribution of flood flows in Tunisia. *Hydrology and Earth System Sciences Discussions, 4(2), 957–981.*

Ahmad, Ishfaq, Fawad, M., Akbar, M., Abbas, A., and Zafar, H. (2016). Regional frequency analysis of annual peak flows in Pakistan using linear combination of order statistics. *Polish Journal of Environmental Studies, 25(6), 2255–2264. https://doi.org/10.15244/pjoes/63782.*

Ahmad, I., Ahmad, T., & Almanjahie, I. M. (2019). Modelling of extreme rainfall in Punjab: Pakistan using Bayesian and frequentist approach. *Applied Ecology and Environmental Research, 17(6), 13729–13748. https://doi.org/10.15666/aeer/1706_1372913748.*

Ahmad, T., Ahmad, I., Arshad, I. A., & Bianco, N. (2021). A comprehensive study on the Bayesian modelling of extreme rainfall: A case study from Pakistan. *International Journal of Climatology*, *42*(1), 208-224. https://doi.org/10.1002/joc.7240

Cunnane, C. (1988). Methods and merits of regional flood frequency analysis. *Journal of Hydrology, 100(1–3), 269–290.* https://doi.org/10.1016/0022-1694(88)90188-6

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer.

Cook, N.J. and Harris, R.I. (2001). Discussion on application of the generalized Pareto distribution to extreme value analysis in wind engineering by JD Holmes, WW

Moriarty. *Journal of Wind Engineering and Industrial Aerodynamics, 89(2), 215–224.*

Campos, R.M. and Soares, C.G. (2016) Regional frequency analysis of wind speed on the coast of Portugal. In: Progress in Renewable Energies Offshore: *Proceedings of the 2nd International Conference on Renewable Energies Offshore (RENEW2016), Lisbon, Portugal. CRC Press, p. 119.*

Darbandi, S., Aalami, M.T. and Asadi, H. (2012). Comparison of four distributions for frequency analysis of wind speed. *Environment and Natural Resources Research, 2(1), 96.* doi**:**10.5539/enrr.v2n1p96

Fowler, H. J., and Kilsby, C. G. (2003). A regional frequency analysis of United Kingdom extreme rainfall from 1961 to 2000. *In International Journal of Climatology 23(11), pp. 1313–1334). https://doi.org/10.1002/joc.943*

Fawad, M., Ahmad, I., Nadeem, F. A., Yan, T., and Abbas, A. (2018). Estimation of wind speed using regional frequency analysis based on linear-moments. *International Journal of Climatology, 38(12), 4431–4444. https://doi.org/10.1002/joc.5678*

Goel, N. K., Burn, D. H., Pandey, M. D., & An, Y. (2004). Wind quantile estimation using a pooled frequency analysis approach. *Journal of wind engineering and industrial aerodynamics*, *92*(6), 509-528. https://doi.org/10.1016/j.jweia.2004.02.003

Ghazijahani, N. and Lee, T. (2015). Mixture distribution fitting and modeling for regional frequency analysis of maximum wind speed over South Korea. *Journal of Korean Society of Hazard Mitigation, 15(5), 191–199.* doi:10.9798/KOSHAM.2015.15.5.191

Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 105-124. https://www.jstor.org/stable/2345653

Hosking, J.R. and Wallis, J.R. (1993) Some statistics useful in regional frequency analysis. *Water Resources Research, 29(2), 271–281*. https://doi.org/10.1029/92WR01980

Hosking, J.R. and Wallis, J.R. (1997) *Regional Frequency Analysis: An Approach based on L-moments*. Cambridge: Cambridge University Press.

Harris, I. (2005) Generalised Pareto methods for wind extremes. Useful tool or mathematical mirage? *Journal of Wind Engineering and Industrial Aerodynamics, 93(5),341–360*.10.1016/j.jweia.2005.02.004

Hussain, Z., and Pasha, G. R. (2009). Regional flood frequency analysis of the seven sites of Punjab, Pakistan, using L-moments. *Water Resources Management, 23(10), 1917–1933*. https://doi.org/10.1007/s11269-008-9360-7.

Hong, H.P., Li, S.H. and Mara, T.G. (2013) Performance of the generalized least-squares method for the Gumbel distribution and its application to annual maximum wind speeds. *Journal of Wind Engineering and Industrial Aerodynamics, 119, 121–132*.

https://doi.org/10.1016/j.jweia.2013.05.012

Hong, H.P. and Ye, W. (2014) Estimating extreme wind speed based on regional frequency analysis. *Structural Safety, 47, 67–77*. https://doi.org/10.1016/j.strusafe.2013.12.001

Komi, K., Amisigo, B. A., Diekkrüger, B., & Hountondji, F. C. (2016). Regional flood frequency analysis in the Volta River basin, West Africa. *Hydrology*, *3*(1), 5. https://doi.org/10.3390/hydrology3010005

Lun, I. Y. F., and Lam, J. C. (2000). A study of Weibull parameters using long-term wind observations. *Renewable Energy, 20(2),* 145–153. https://doi.org/10.1016/S0960-1481(99)00103-2

Modarres, R. (2008) Regional maximum wind speed frequency analysis for the arid and semi-arid regions of Iran. *Journal of Arid Environments, 72(7), 1329–1342*.

https://doi.org/10.1016/j.jaridenv.2007.12.010

Nedaei, M., Assareh, E., & Biglari, M. (2014). An extensive evaluation of wind resource using new methods and strategies for development and utilizing wind power in Mah-shahr station in Iran. *Energy conversion and management*, *81*, 475-503.

https://doi.org/10.1016/j.enconman.2014.02.025.

Naghettini, M. (Ed.). (2017). *Fundamentals of statistical hydrology*. Cham: Springer International Publishing.

Liu, D., Wang, L., & Pang, L. (2006). Theory of multivariate compound extreme value distribution and its application to extreme sea state prediction. *Chinese Science Bulletin*, *51*(23), 2926-2930. https://doi.org/10.1007/s11434-006-2186-x.

Lombardo, F. T., Main, J. A., & Simiu, E. (2009). Automated extraction and classification of thunderstorm and non-thunderstorm wind data for extreme-value analysis. *Journal of Wind Engineering and Industrial Aerodynamics*, *97*(3-4), 120-131. https://doi.org/10.1016/j.jweia.2009.03.001.

Lee, B. H., Ahn, D. J., Kim, H. G., & Ha, Y. C. (2012). An estimation of the extreme wind speed using the Korea wind map. *Renewable energy*, *42*, 4-10. https://doi.org/10.1016/j.renene.2011.09.033.

Liang, B., Shao, Z., Li, H., Shao, M., & Lee, D. (2019). An automated threshold selection method based on the characteristic of extrapolated significant wave heights. *Coastal Engineering*, *144*, 22-32. https://doi.org/10.1016/j.coastaleng.2018.12.001.

Ohunakin, O. S., & Akinnawonu, O. O. (2012). Assessment of wind energy potential and the economics of wind power generation in Jos, Plateau State, Nigeria. *Energy for Sustainable Development*, *16*(1), 78-83. https://doi.org/10.1016/j.esd.2011.10.004.

Parida, B. P., Kachroo, R. K., and Shrestha, D. B. (1998). Regional Flood Frequency Analysis of Mahi-Sabarmati Basin (Subzone 3-a) using Index Flood Procedure with L-Moments. *Water Resources Management, 12(1), 1–12*. https://doi.org/10.1023/A:1007970800408

Randriamihamison, N., Vialaneix, N., & Neuvial, P. (2021). Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *Journal of Classification*, *38*(2), 363-389.

Shiau, J.T. and Wu, F.C. (2009) Regionalization of natural flow regime: application to environmental flow optimization at ungauged sites. *River Research and Applications, 25(9), 1071–1089*.

Seckin, N., Haktanir, T. and Yurtal, R. (2011) Flood frequency analysis of Turkey using L-moments method. *Hydrological Processes, 25(22), 3499–3505*. doi:10.1002/hyp.8077

Shahzadi, A. (2013) A review: regional frequency analysis of annual maximum rainfall in monsoon region of Pakistan using l-moments. *International Journal of Advanced Statistics and Probability*, 1(3), 97–101. doi: 10.1234/pjsor.v9i1.461

Tizpar, A., Satkin, M., Roshan, M. B., & Armoudli, Y. (2014). Wind resource assessment and wind power potential of Mil-E Nader region in Sistan and Baluchestan Province, Iran–Part 1: Annual energy estimation. *Energy conversion and management*, *79*, 273-280.

https://doi.org/10.1016/j.enconman.2013.10.004.

Weisser, D. (2003). A wind energy analysis of Grenada: an estimation using the 'Weibull'density function. *Renewable energy*, *28*(11), 1803-1812.

https://doi.org/10.1016/S0960-1481(03)00016-8.

Wang, J., Qin, S., Jin, S., & Wu, J. (2015). Estimation methods review and analysis of offshore extreme wind speeds and wind energy resources. *Renewable and Sustainable Energy Reviews*, *42*, 26-42. https://doi.org/10.1016/j.rser.2014.09.042.

Yue, S. and Wang, C. (2004) Determination of regional probability distributions of Canadian flood flows using L-moments. *Journal of Hydrology (New Zealand), 1, 59–73.*

https://www.jstor.org/stable/43944914

Yao, Z., Xiao, J., & Jiang, F. (2012). Characteristics of daily extreme-wind gusts along the Lanxin Railway in Xinjiang, China. *Aeolian Research*, *6*, 31-40. https://doi.org/10.1016/j.aeolia.2012.07.002.

Figures

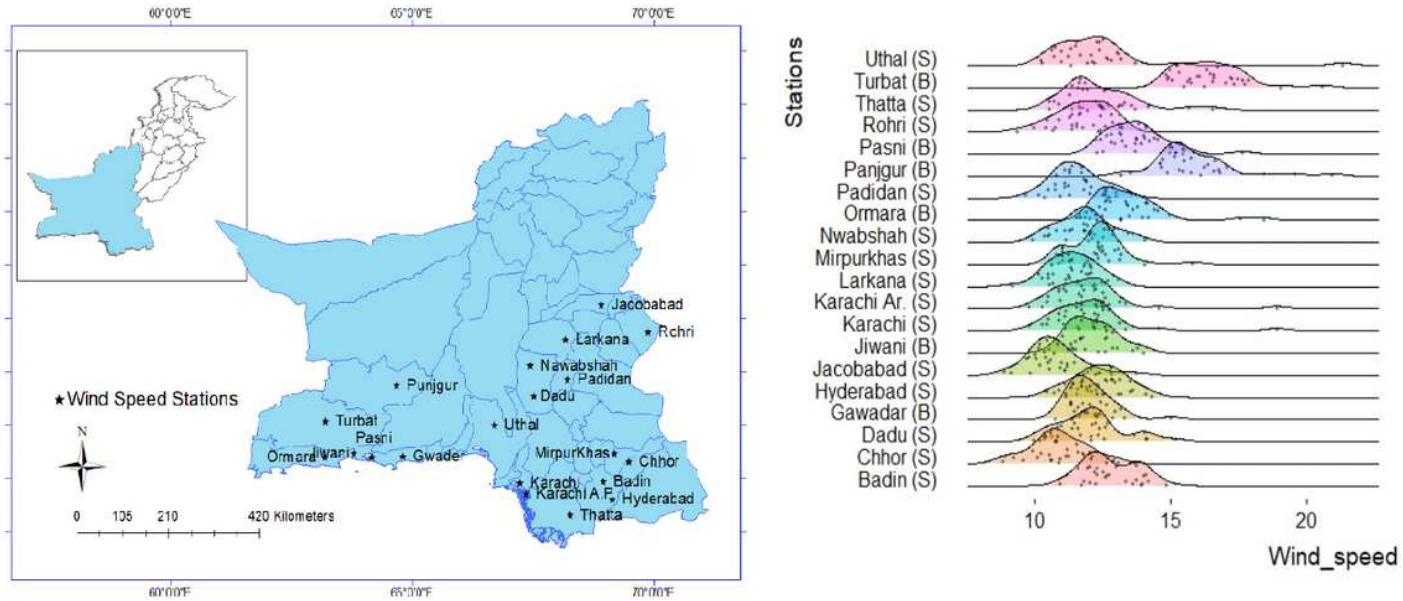

Figure 1

Left: Geographical existence of the stations used in this study. Right: Ridgeline plots of daily annual maximum wind data for all stations with climatic regions Sindh and Baluchistan.

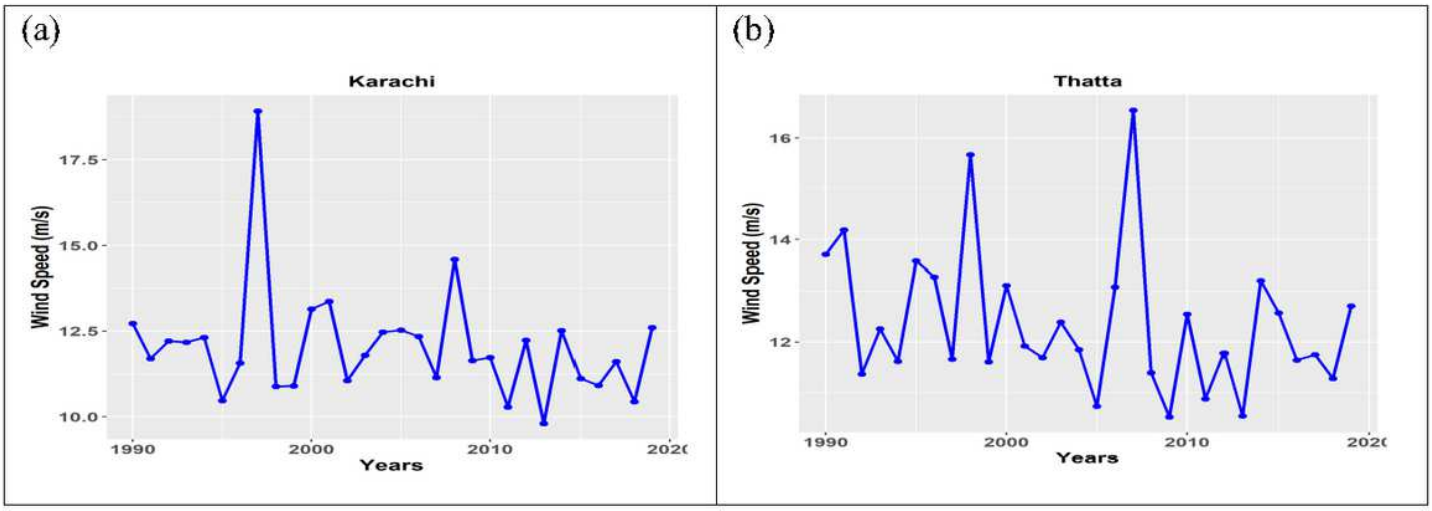

Figure 2

Historigrams of Karachi and Thatta stations.

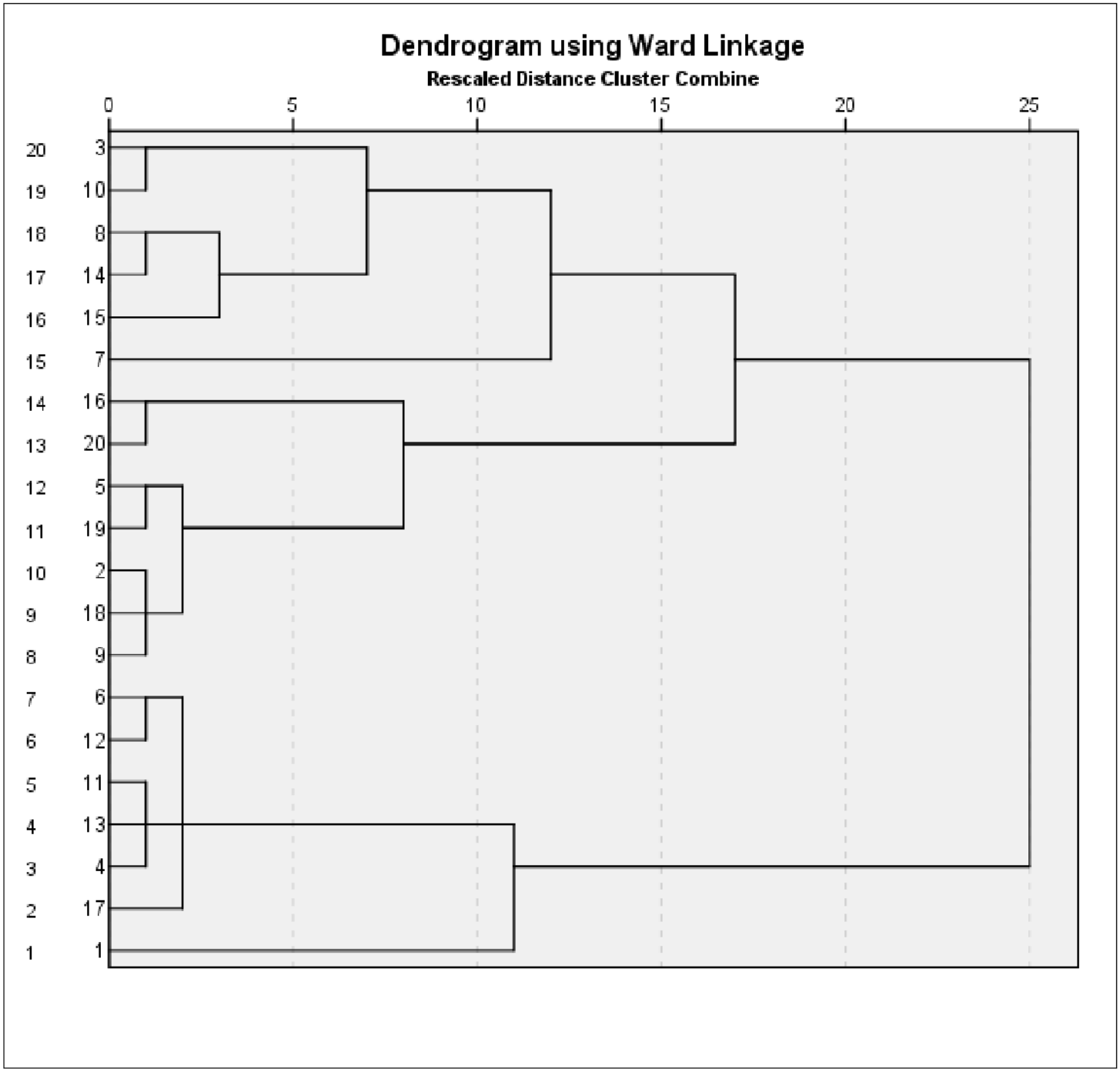

Figure 3

Dendrogram for Cluster I and II.

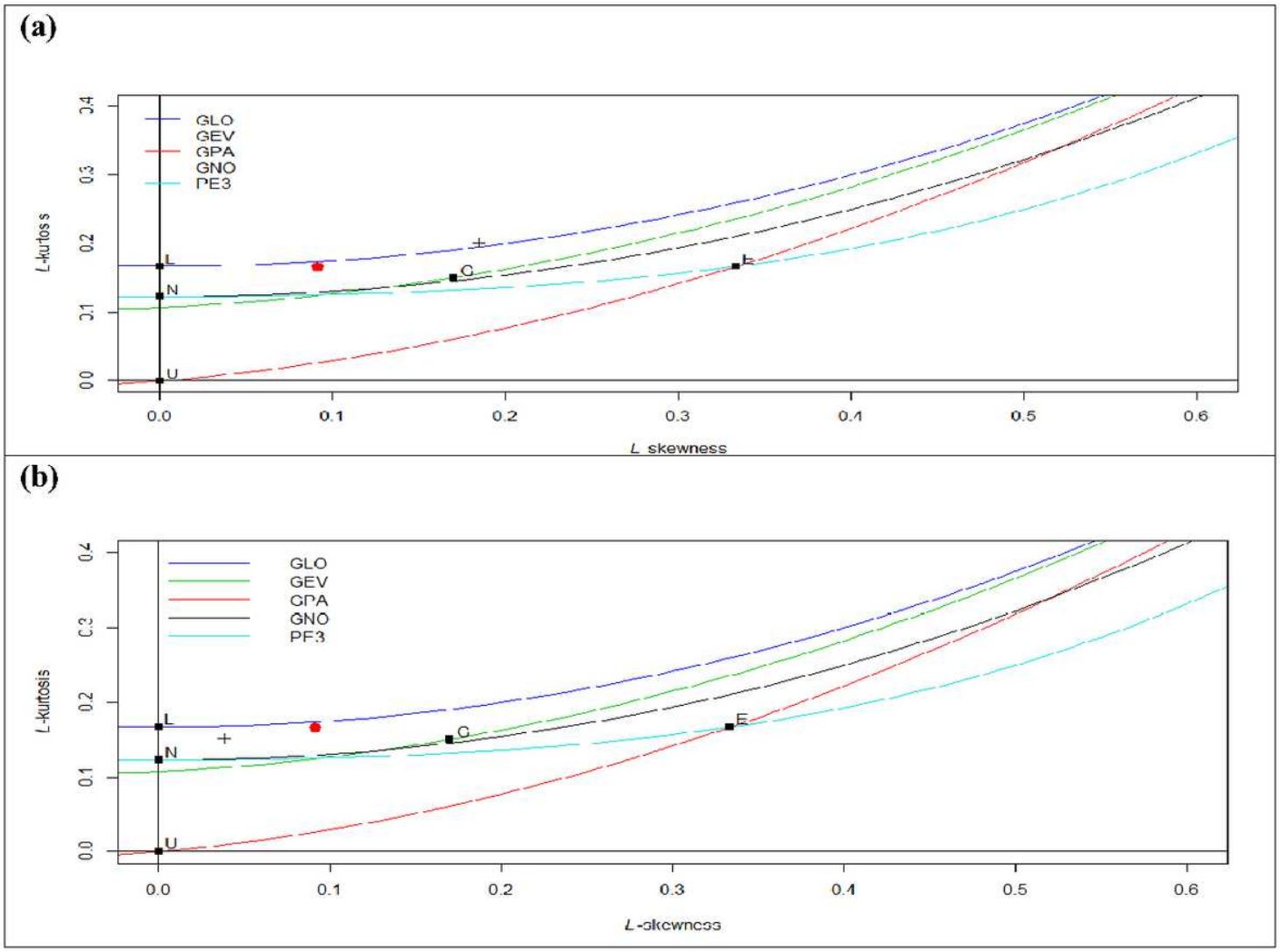

Figure 4

(a) L-moment ratio diagram for Region I (b) L-moment ratio diagram for Region II

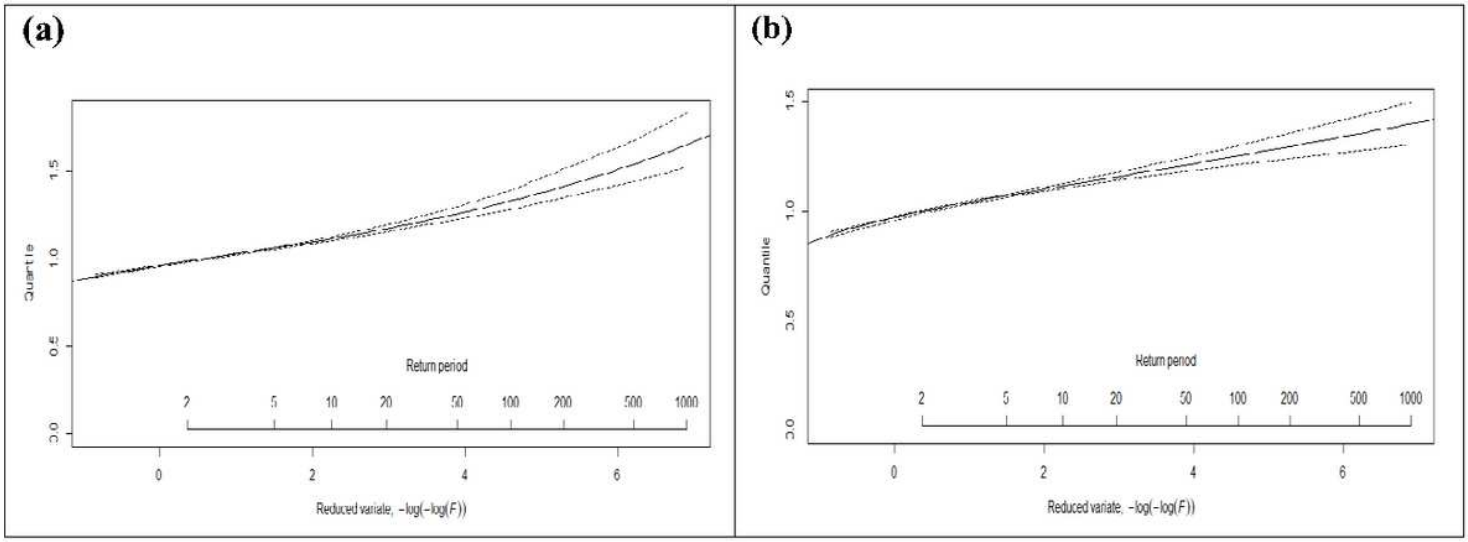

Figure 5

(a) GLO distribution regional growth curve for Region I (b) GLO distribution regional growth curve for Region II.

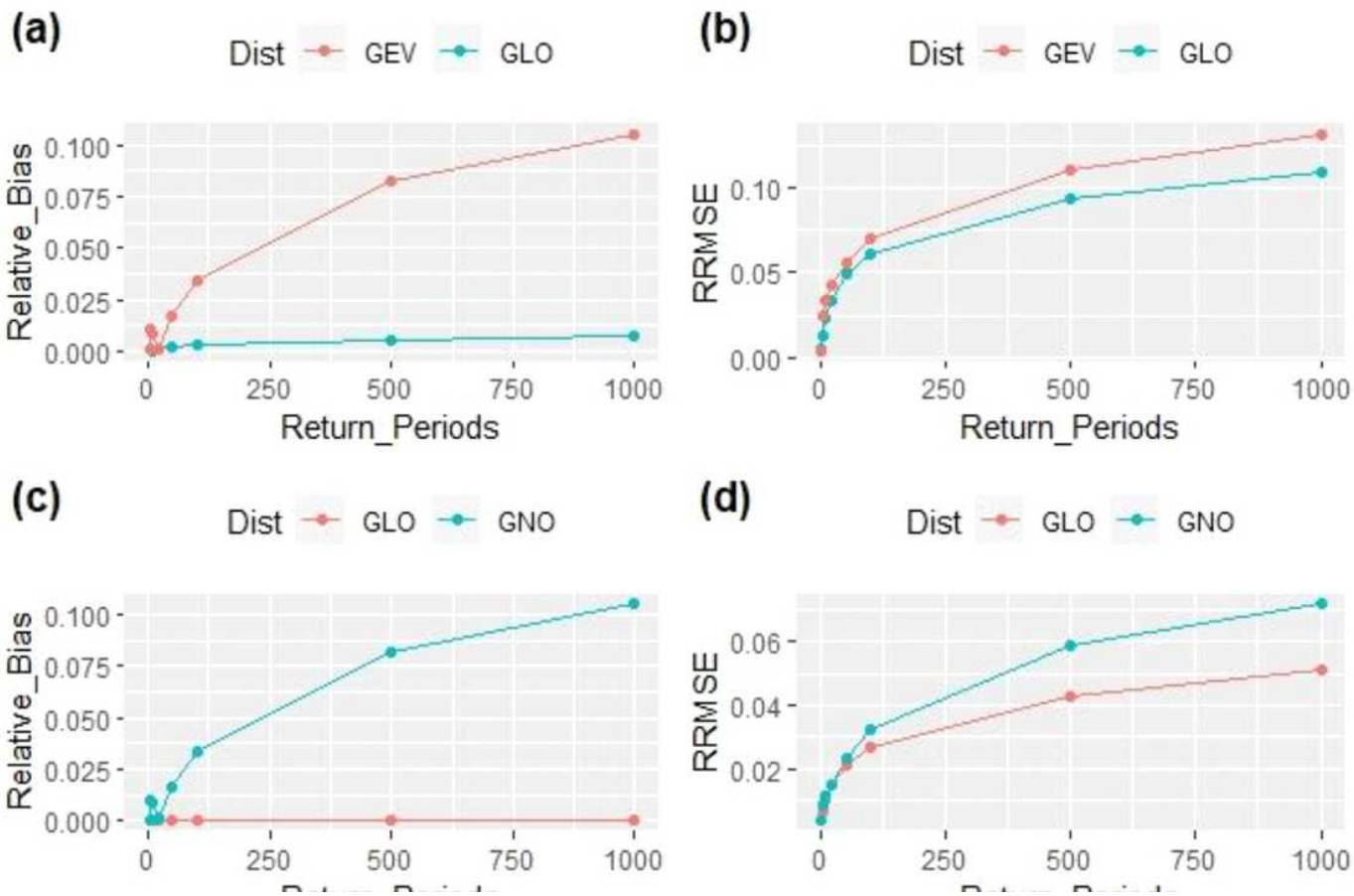

Figure 6

Graphical display of assessment measures for best-fit distributions. (a, b) Relative bias and relative root mean square error for GLO and GEV distribution corresponding to region I, (c, d) Relative bias and relative root mean square error for GLO and GNO distribution corresponding to region II.