Comment on "Biases in Estimating Long-Term Recurrence Intervals of Extreme Events Due To Regionalized Sampling" by El Rafei et al. (2023)

Chi-Hsiang Wang[1] and John D. Holmes[2]

[1]Corresponding author: Chi-Hsiang Wang (chi-hsiang.wang@csiro.au), CSIRO Energy, Clayton, Victoria, Australia, [2]JDH Consulting, Mentone, Victoria, Australia

Key Points:

- Grouping data properly de-trended from stations of consistent meteorology does not induce biases at long recurrence intervals
- The 'superstation' technique reduces the sampling errors of short-length data from individual stations of consistent meteorology
- Consistent meteorology is required for grouping stations corrected for non-standard terrain and effects of local topography

Abstract

The 'super-station' approach has been adopted since 1980s as a pragmatic method of improving extreme-value predictions by grouping short-length datasets from several measurement stations to become a larger dataset to reduce uncertainties due to random sampling variation. El Rafei *et al.* (2023, https://doi.org/10.1029/2023GL105286) analyzed reanalysis, and randomly generated, wind extremes datasets and suggested that this technique can introduce unexpected biases in typical situations. We complement their work and demonstrate by Monte-Carlo simulation, assuming the same number of grouped stations and data lengths used, that applying the grouping technique to samples of properly de-trended datasets to meet the homogeneity assumption does *not* lead to biased prediction of extremes. In addition, the grouping technique effectively reduces the uncertainty and sampling errors that result from short-length datasets from individual stations of consistent meteorology.

Plain Language Summary

Pooling extremal data observed from different sites of consistent environment for analysis and treating the pooled data as if they were observed at one site has been in practice for 40 years. A recent study reckoned such data pooling introduces bias errors in typical situations. We repeat their analysis by random-number generation plus a data homogenizing step and show that the data-pooling technique does not cause bias errors. Instead, the technique is effective in reducing the random errors experienced when analyzing an un-pooled small dataset.

1 Introduction

Complementing El Rafei et al (2023), the main objective of this comment is to clarify that the superstation approach is an unbiased approach for analysis of grouped data sets *that conform to consistent meteorology*. The superstation technique was developed based on the assumption of homogeneous meteorology (Dorman 1983a; Palutikof *et al*. 1999) in an effort to reduce statistical sampling errors. To fulfil the assumed homogeneity, the data sets to be grouped as a superstation need to be ‘de-trended’ (i.e. ‘homogenized’) (Peterka, 1992; Palutikof *et al*., 1999; Torrielli *et al*., 2013). The de-trending step appears not carried out in El Rafei et al (2023), which leads to the bias the authors suggested as being introduced by the superstation approach. To facilitate fulfilment of the homogeneity assumption, we present a de-trending algorithm that produces a superstation representative of the ‘average’ behaviour of a consistent environment, and demonstrate the de-trending step by reproducing the result in Figure 2A of El Rafei *et al*. (2023).

2 Superstation technique for homogeneous meteorology

El Rafei *et al.* (2023) have used convective wind gusts from Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia Sydney region (BARRA-SY) and from Monte Carlo simulation of heterogeneous probability distributions, as examples of wind gust hazard analysis using the superstation approach, and suggested that the grouping approach can introduce unexpected biases in typical situations at long recurrence intervals. Thus, in their ‘Conclusion’ section, for example, the authors state: “*The superstation fit tends to the highest levels suggested by any of the pooled locations and this bias increases with longer recurrence intervals*”. We intend to emphasize that the suggested bias occurs, unsurprisingly, to situations where data sets are grouped as a superstation, but not analyzed by a method that conforms to the underlying assumption of meteorological homogeneity. On the other hand, if meteorological homogeneity is fulfilled, this data-grouping technique is unbiased and effectively reduces the uncertainty and sampling errors that result from short-length datasets of individual stations.

It has been well recognized that the superstation approach “… is only feasible in a climatologically homogeneous area” (Palutikof *et al*. 1999). Furthermore, when homogeneity is assumed, the data sets to be grouped need to be ‘de-trended’, as pointed out by Palutikof *et al.* (1999), “… de-trending was carried out using the mean annual extreme, so that the resulting superstation data may be expected to be homogeneous in the upper tail.”

An apparent difficulty faced by the analysts is a lack of criteria for gauging the appropriateness of data grouping. Even though statistical techniques have been attempted (e.g. Dorman 1983b) for this purpose, to our knowledge, no clear criteria based on physical reasoning have been developed. Under such circumstances, the analysts must exercise judgments on whether homogeneity is acceptable for their objective. If they decide to regard the data sets as being from consistent meteorology and apply the superstation technique, then dataset de-trending should be performed before conducting extreme-value analysis. On the other hand, if the data sets are regarded, or known, to be from heterogeneous underlying statistics, the superstation technique is not applicable such that they should be analyzed separately.

El Rafei *et al.* rightly noted that: *“… These superstations represent specific geographical areas where stations with meteorological consistency are grouped together…*” However, the results from their analyses appear not to follow the consistent meteorological assumption or not to analyze the data sets properly. For instance, Figure 2A presents a superstation derived by grouping five data sets, each was generated by one of five generalized Pareto distributions (GPDs). The application of superstation technique implies that the data sets were regarded to be from consistent meteorology. If this is indeed the authors’ intention, the de-trending step should be performed. However, this appears not to be the case, as judged from the resulting superstation which exhibits bias towards the most hazardous model.

In the following, we first show by Monte-Carlo simulation that applying the grouping technique to samples from *homogeneous* datasets does *not* lead to biased prediction of extremes. We then present a de-trending algorithm that facilitates fulfilment of meteorological homogeneity and demonstrate the de-trending step for the resulting superstation illustrated in Figure 2A of El Rafei *et al.* (2023).

3 Validity of the superstation approach

Because of random sampling variation, the extent of uncertainty for estimating the underlying statistical distribution depends on the length of the dataset. This manifests as the extent of uncertainty in the estimated distribution parameters: the longer the dataset, the narrower the confidence intervals (CI's) of the distribution parameters. For example, as illustrated in Figure 1 below, individual station data lengths of 23 years and 1,000 years, as used for the results shown respectively in Figures 1 and 2 in El Rafei *et al.* (2023), lead to different conclusions about whether a specific distribution is accepted as the underlying model of the data.

Five GPD models were used in Figure 2A by El Rafei *et al*. These had the same exceedance rate ($\lambda = 5$), threshold ($u_0 = 20$ m/s) and shape parameter ($\xi = -0.1$), but different scale parameters ($\sigma_i = 2.75 + 0.25i, i = 1,2,\ldots,5$).

To check of validity of the biases by the super-station approach claimed by El Rafei *et al*, their third GPD model (i.e. with $\sigma_3 = 3.5$) is used here to generate synthetic data. We follow the treatment of their paper to generate by Monte-Carlo simulation 23 years of data for 25 hypothetical stations. Figure 1a shows the generated data (thin black lines) and the resulting super-station data (red circular points) along with the five theoretical GPD models (thick colored lines) in the wind gust versus log-ARI plot. Similarly, because Figure 2A of El Rafei *et al.* used 1,000 years of data to obtain the super-station data, we have generated 1,000 years of data for 25 hypothetical stations, as shown in Figure 1b.

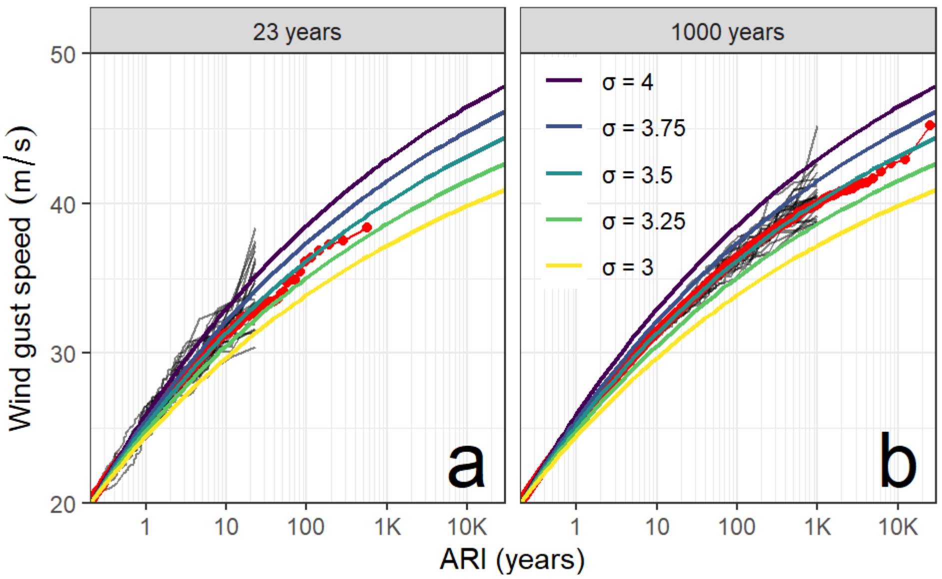

Figure 1: Simulated gust data of 25 hypothetical stations and super-station for (a) 23 years; and (b) 1,000 years.

Figure 1 clearly reveals that, for wind gusts given an ARI, the dataset of 23 years spreads much more widely than the dataset of 1,000 years. This is a result of shorter records being more seriously affected by sampling variation than of longer records. The spread of the 23-year data tracks of the 25 stations covers essentially all the theoretical gust speed values of the five models. That is, given a sample of 23-year data from any individual station, one cannot assert with high confidence which of the five models is the underlying model. On the contrary, with 1,000-year data from an individual station, in the overwhelming cases one is able to deduce with sufficient confidence the third model is the model which generates the dataset. In addition, the super-stations (red circular points) shown in the two cases do not exhibit a systematic tendency of biases towards more hazardous models, as suggested by El Rafei *et al*.

To see more closely the uncertainties in σ_3, 10,000 stations are generated for datasets of 23 and 1,000 years. They have been fitted to the GPD model with the shape parameter being the only unknown. The probability densities of the estimated σ_3 are shown in Figure 2, in which the thick and thin red lines represent 67 % and 95 % CIs, respectively. Figure 2a shows that the 95 % CI for σ_3 for 23-year data from one station is [2.93, 4.08], covering all the shape-parameter values (ranging from 3 to 4) of the five models. This implies that, with 23 years of data in one station, we fail to reject that any of the five models could be the true model. In contrast, the 95 % CI of σ_3 for 1,000 years data from one station is [3.41, 3.59] (Figure 2c), which establishes, with statistical significance, that the third model is the true one.

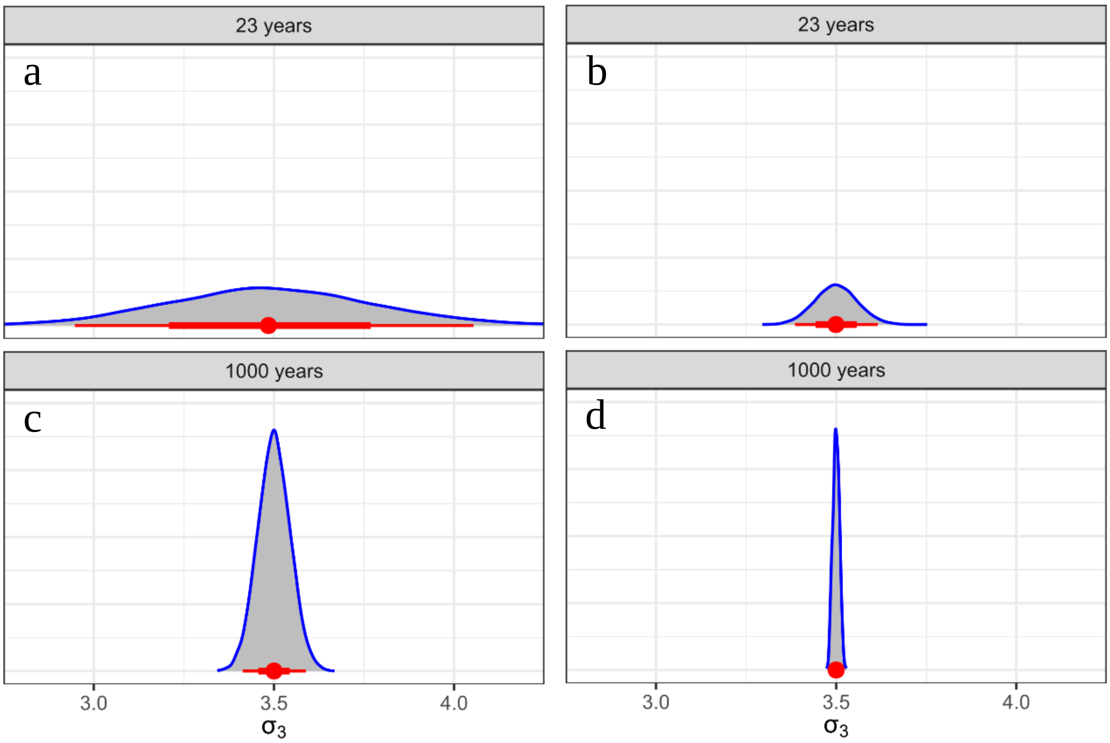

Figure 2: Probability densities and confidence intervals of σ_3 for datasets of (a) 23 years at one station; (b) 23 years at 25 grouped stations; (c) 1,000 years at one station; and (d) 1,000 years at 25 grouped stations.

Comparing Figures 2a and c to Figures 2b and d (produced by grouping data from the 25 hypothetical stations to form super-stations), respectively, illustrates the advantageous effect by data grouping in reducing the variance of σ_3, which is also implied in Figure 1. Importantly, all the point estimates (red circles) do not show biases for the true value of σ_3 due to data grouping.

Another implication of Figure 2 is that the 1,000-year datasets generated by the five different models, as done in El Rafei *et al.* (2023), would indicate clearly that they are generated by five distinct models, which mean indeed the five datasets are from heterogeneous meteorology. Grouping the five datasets into a super-station would violate the basic requirement that they are recorded in regions of consistent meteorology, which is the same basic requirement for estimating the common inferential statistics (e.g. mean and standard deviation) of a dataset drawn from a defined sample space. Therefore, the claimed biases observed in Figure 2A of the paper by El Rafei *et al.* arise from treating datasets from obviously different sample spaces as if they were drawn from one sample space, but not biases due to the application of super-station approach. This also indicates the importance of clearly identifying the sample space of the subject-matter problem before conducting a proper statistical analysis.

4 A 'de-trending' algorithm for superstation approach

Based on the concept of standardizing random variables, this section presents an algorithm that produces the 'average' behavior of homogenized meteorology manifested as the de-trended superstation.

Suppose that there are *J* stations and the *j*-th station provides N_j wind speeds, i.e. $v_{i,j}$, where $i = 1, \ldots, N_j$. Then

1. For the *j*-th station data:
   i) Estimate the mean, $\bar{v}_j = \sum_{i=1}^{N_j} v_{i,j} / N_j$, and standard deviation, $s_{v,j} = \sqrt{\sum_{i=1}^{N_j} \left(v_{i,j} - \bar{v}_j\right)^2 / \left(N_j - 1\right)}$, of $v_{i,j}$'s.
   ii) Compute the standardized wind speeds, $v'_{i,j} = \left(v_{i,j} - \bar{v}_j\right) / s_{v,j}$
2. The standardized wind speeds from the *J* stations are then grouped and transformed:
   i) Estimate the grand mean, $\bar{v}_g = \sum_{j=1}^{J} \bar{v}_j / J$ and grand standard deviation $s_g = \sum_{j=1}^{J} s_{v,j} / J$, as the mean and standard deviation of the wind-speed probability distribution which may be regarded as the 'mean' underlying wind-generating mechanism.
   ii) Group the *J*-station data to form a standardized superstation data, v'_k, $k = 1, \ldots, K$, where $K = \sum_{j=1}^{J} N_j$.
   iii) Compute the superstation wind speeds $v_k = \bar{v}_g + s_g v'_k$, which are used to analyze the hazard of the homogenized region.

It should be emphasized that such 'homogenization' process should be done to data sets that the analysts regard as being collected from homogeneous meteorology. As an example to illustrate the de-trending step for the case presented in Figure 2A of El Rafei *et al.* (2023), we used each model in that figure to generate one data set of 50,000 years of gust speeds, for a total of five data sets which was then de-trended. This leads to a superstation with 250,000 station-years. Figure 3 shows the results. Up to 10,000-year ARI (the highest ARI in Figure 2A of El Rafei *et al.*), the thick non-de-trended red line is close to the superstation represented by square points in Figure 2A of El Rafei *et al.*, (2023), while the thick de-trended blue line (representing the average behavior of assumed homogeneous meteorology) happens to follow closely the model with $\sigma = 3.5$.

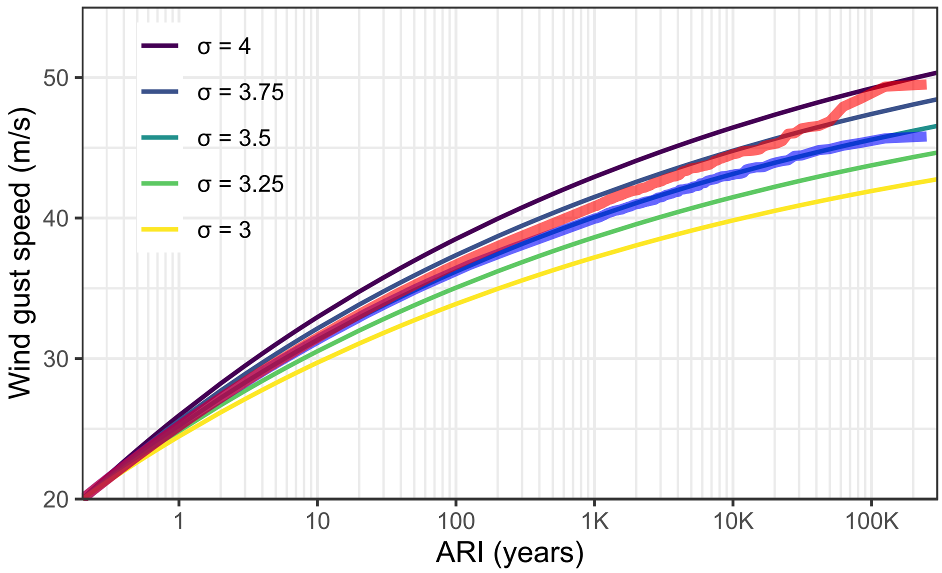

Figure 3: Simulated superstation data that are from heterogeneous (red line) and 'homogenized' (blue line) environments

5 Correction for local terrain and topographical effects and the GPD-GEV duality

El Rafei *et al.* did not use recorded and corrected surface wind data, but instead distributions from BARRA-SY reanalysis. When processing recorded anemometer data from surface weather stations, correction for terrain to standard conditions can be as high as 20 % with topographic corrections even greater (e.g. Holmes 2016; Australian Standards 2021); however it is not clear how this is done with the reanalysis-derived gusts. Secondly the data set only extends to 23 years (1996 to 2019). The 'speckled' values in Figure 1A derived from such a short period therefore may contain significant sampling errors.

El Rafei *et al.* (2023) also observed at the end of Section 3 that: "*Both (GPD and GEV) show a similar level of bias for all record lengths, although the biases are slightly smaller if GPD is used instead of the usual (for rainfall) GEV.*" The GPD and GEV have been known to possess a duality relationship: for a given GPD model, an equivalent GEV model can be found, and vice versa (Wang and Holmes, 2020). That is, if the same post-processed dataset is analyzed, both GPD and GEV should give exactly the same result because of the duality relationship, e.g. for the model with $\sigma_3 = 3.5$ in Figure 2A of El Rafei et al (2023), the parameters of its equivalent GEV are $u_{0_g} = 25.2$, $\sigma_g = 2.98$, and $\xi_g = -0.1$. It is not clear how the data sets for the rainfall example were processed for fitting to the two probability models. For example, processing the raw data sets into a block-maxima and a peaks-over-threshold data sets would invariably produce two different data sets that lead to two different fitted models. Therefore, the observed performance difference between the GPD and GEV may originate from data processing or model fitting step rather than inherently disparate properties of the models.

6 Summary

We have shown by simulation of samples from the *same underlying probability distribution*, i.e. *homogeneous* datasets, that the grouping technique does *not* lead to biased prediction of extremes, as previously suggested by El Rafei *et al* (2023). We have presented a de-trending algorithm for homogenizing the data sets to be grouped as a superstation to meet its homogeneity assumption. Moreover, the superstation technique is shown to reduce the uncertainty and sampling errors resulting from prediction from datasets from individual stations of short length, provided that datasets from similar climates are grouped, and that they are corrected for non-standard terrain and for any effects of local topography.

References

Australian / New Zealand Standard. (2021). "Structural design actions Part 2: Wind actions."

Dorman C.M.L. (1983a). Extreme wind gusts in Australia, excluding tropical cyclones. *Civil Engineering Transactions*, Institution of Engineers, Australia, CE25, 96-106.

Dorman, C.M.L. (1983b). United States extreme wind speeds — a new view. *Journal of Wind Engineering and Industrial Aerodynamics*, 13, 105-114. https://doi.org/10.1016/0167-6105(83)90133-2

El Rafei, M., Sherwood, S., Evans, J., Dowdy, A. and Ji, F. (2023). Biases in estimating long-term recurrence intervals of extreme events due to regionalized sampling. *Geophysical Research Letters* 50 (15): e2023GL105286. https://doi.org/10.1029/2023GL105286

Holmes, J.D. (2016). *Determination of turbulence intensity and roughness length from AWS data*. Paper presented at 18th Australasian Wind Engineering Society Workshop, McLaren Vale, South Australia.

Palutikof, J. P., B. B. Brabson, D. H. Lister, and S. T. Adcock. (1999). "A review of methods to calculate extreme wind speeds." Meteorological Applications 6 (2): 119–32. https://doi.org/10.1017/S1350482799001103.

Peterka, J. A. (1992). "Improved extreme wind prediction for the United States." Journal of Wind Engineering and Industrial Aerodynamics 41 (1-3): 533–41. https://doi.org/10.1016/0167-6105(92)90459-N.

Torrielli, A., M. P. Repetto, and G. Solari. (2013). "Extreme wind speeds from long-term synthetic records." Journal of Wind Engineering and Industrial Aerodynamics 115: 22–38. https://doi.org/10.1016/j.jweia.2012.12.008.

Wang, C-H. and Holmes, J.D. (2020). Exceedance rate, exceedance probability and the duality of GEV and GPD for extreme hazard analysis. *Natural Hazards*, Vol. 102, pp 1305-1321, https://doi.org/10.1007/s11069-020-03968-z