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 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 claimed that this technique can introduce unexpected biases in typical situations. We demonstrate by Monte-Carlo simulation, assuming the same number of grouped stations and data lengths used, that applying the grouping technique to samples from *homogeneous* datasets 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 and found the data-pooling technique does not cause bias errors. Instead, the technique is effective in reducing the random errors experienced when analyzing an unpooled small dataset.

1 Introduction

This contribution is concerned with the common practice of grouping recorded extreme climatic variables from measurement stations with perceived similar climates since 1980s (Dorman 1983), in an effort to reduce statistical sampling errors from short record lengths, when making predictions for low probabilities of exceedance, or high average recurrence intervals (ARIs). We use Monte-Carlo simulated extreme wind speeds to illustrate the approach, as employed by El Rafei *et al* (2023).

2 Comments on El Rafei *et al* (2023)

El Rafei *et al* (2023) have used convective wind gusts in New South Wales in Australia as examples and concluded that the grouping approach leads to biases in estimates of geophysical variables with high ARIs. Thus, in their 'Conclusion' section, 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*". However, several questions come to mind with respect to the analyses by El Rafei *et al*.

The authors have not used recorded and corrected surface wind data, but instead distributions from BARRA-SY reanalysis. When processing recorded anemometer data from surface weather stations, it is necessary to correct for terrain and topography (e.g. Holmes, 2016); however it is not clear how this is done with the reanalysis-derived gusts. Are uncorrected terrain and topographic effects a reason for the 'speckling' in Figure 1A?

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 themselves contain significant sampling errors.

Thirdly, the 50 km by 50 km group size is relatively small. Although individual convective downburst events are smaller, they often occur sequentially in multiple cells (e.g. 'squall lines') that affect much larger areas. A good example is the event in South Australia in September 2016 that caused failures of several transmission lines (Australian Bureau of Meteorology 2016).

The simulation analysis leading to Figure 2 of their paper is also puzzling. The authors have deliberately varied the parameters of the underlying probability distributions when generating synthetic data. Hence, the combined data is heterogeneous. Is it then legitimate to make conclusions about a grouping method that assumes homogeneity, using inhomogeneous synthetic data?

El Rafei *et al* noted that: *"... These superstations represent specific geographical areas where stations with meteorological consistency are grouped together...,*". Therefore, all the locations where the datasets are recorded should be ensured as having one and the same underlying statistical distribution which randomly generates the data points grouped.

Because of random sampling variation, however, 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 the following section, 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 Generalized Pareto distributions (GPD) 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 ($\sigma_i = 2.75 + 0.25i, i = 1,2,\ldots,5$).

El Rafei *et al* (2023) also claimed, at the end of Section 3 of that paper, 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, it is unnecessary to reprocess a block-maxima dataset into a peaks-over-threshold. For example, for the model with $\sigma_3 = 3.5$ in Figure 2A of their paper, the parameters of its equivalent GEV are $u_{0_g} = 25.2$, $\sigma_g = 2.98$, and $\xi_g = -0.1$.

3 Validity check of the grouping approach by simulations

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 the paper by 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.

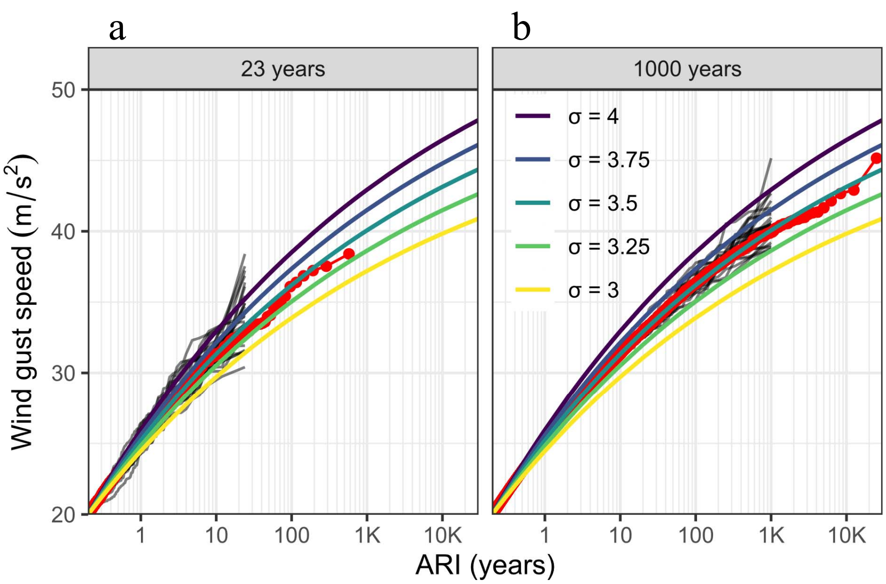

Figure 1: Simulated gust data of 25 hypothetical stations and super-station for (a) 23 years; and (b) 1,000 years.

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 manifest 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 claimed by the authors.

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.

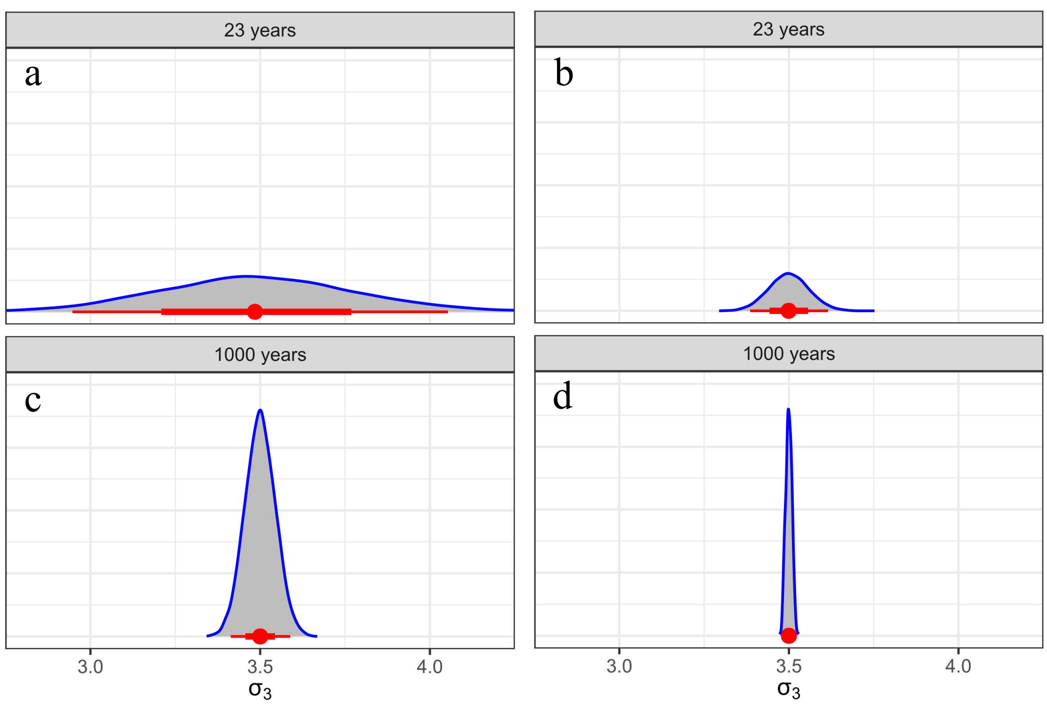

Figure 2: Probability densities and confidence intervals of σ_3 for datasets of (a) 23 years at one station; (b) 23 years at 25 grouped stations; (c) 1,000 years at one station; and (d) 1,000 years at 25 grouped stations.

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 subject-matter problem before conducting a proper statistical analysis.

4 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 claimed by El Rafei *et al* (2023). However, the 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 Bureau of Meteorology (2016). *Severe thunderstorm and tornado outbreak South Australia 28 September 2016*. Retrieved from https://www.dpc.sa.gov.au/__data/assets/pdf_file/0007/15199/Attachment-3-BoM-Severe-Thunderstorm-and-Tornado-Outbreak-28-September-2016.pdf.

Dorman, C. (1983). 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.

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