Nasa SpaceApps challenge

Dr. Thomai Tsiftsi, Dr James P. Edwards, Dr Gourouni Ph.D.,
Det Sup. Leodaris Edwards, H Mouma Edwards

May 1, 2017

In this challenge we examine the Hi-SEAS solar radiation data and employ extreme value analysis to describe the probabilities of extreme values. In particular, we focus on asymptotic analysis of the distributions of particularly low total daily radiation output of the solar panels. The motivation of this analysis is to facilitate the Hi-SEAS crew planning for the possibility of extreme drops in solar panel output (due to natural variation in weather, radiation output and other factors), ensuring that sufficient power reserves are maintained to cope with extreme lows.

The analyse our data we began by estimating the total daily output by approximating the total energy generated by the cells as a discretisation of the time integral of the flux power,

$$x_i = \sum_{k=1}^{N} 72 \times p_{ik} \times \frac{t_{i,k+1} - t_{i,k}}{3600} \times \mathcal{E} \tag{1}$$

where $\mathcal{E} \approx 0.1$ is the estimated efficiency of the solar cells which have area $72m^2$. The sum is over the given data points, p_{ik}, representing the solar radiation flux on day i at time $t_{i,k}$. In anticipation of extreme value analysis, we also blocked these data into weekly observations, $\{x_1, \ldots, x_{n=7}\}$. These data are displayed below in figure (1).

Generalised extreme value theory

Let $M_i = \max\{x_1, \ldots, x_n\}$. Then extreme value analysis states that the asymptotic distribution of M_n, if it exists, must take the form (GEV)

$$\mathcal{P}(M_i \leq z) \simeq G(z|\mu, \sigma, \xi) \equiv \exp\left[-\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right)\right], \tag{2}$$

for parameters μ, location, σ, scale, and shape, ξ. These parameters can be fitted to our data, following which $G(z)$ provides an asymptotic estimate of the tails of the distribution of the maxima of a set of data. Here we are interested in *minima* so we must take our daily accumulated energies and *negate* them so that minima become (negative) maxima.

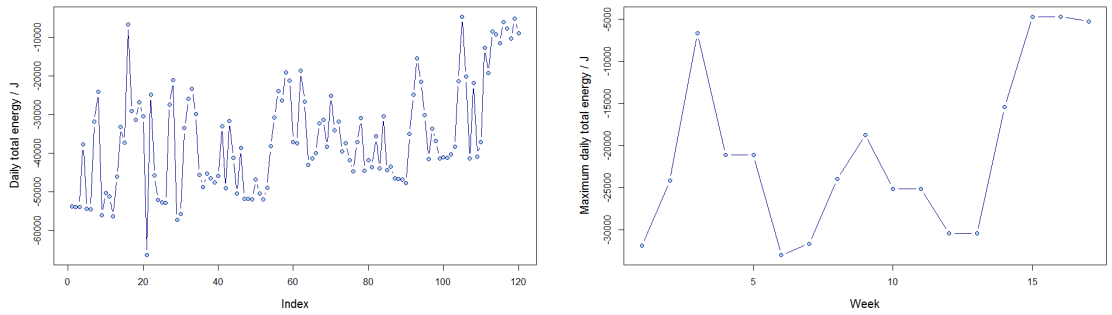
Fitting our data to the model initially provided *maximum likelihood*[1] estimates of the parameters

	μ	σ	ξ
Estimate	-2.6×10^4	48.3×10^3	3.9×10^{-2}
Std Err	2.3×10^3	2.8×10^3	4.6×10^{-1}

[1]The log likelihood for the data given the model is

$$-N\log(\sigma) - (1 + \frac{1}{\xi})\sum_{i=1}^{N}\left[\log\left(1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right) - \left(1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right] \tag{3}$$

where z_i are the empirical maxima. The case $\xi = 0$ is treated as the limiting form of $G(z|\mu, \sigma, \xi \to 0)$. The likelihood is maximised with respect to the parameters μ, σ, ξ.



(a) The estimated total daily energy output by Hi-SEAS polar panels for 120 days.

(b) The maximum daily energy output in each week during the same period.

Figure 1: Daily maxima and the weekly maxima energy output. The energy output has been negated in anticipation of our extreme value analysis, so extreme lows now appear as the largest values.

However, one notes that the shape parameter has large standard errors making its estimate compatible with 0. For this reason we re-fit the data to the subset of models with vanishing shape parameter (Gumbel distribution); the likelihood ratio test provided a p-value of 0.89 on a null hypothesis of $\xi = 0$, giving sufficient evidence to accept it. This lead to a refined model with fewer parameters, whose estimates are

	μ	σ
Estimate	-2.6×10^4	8.3×10^3
Std Err	2.1×10^3	1.8×10^3

which are used as estimates for prediction of extreme daily energy outputs per week.

The plot in figure 2 gives diagnostic and predictive results for the data. The probability distribution function at at z gives the probability that M_n will be within an infinitesimal interval, dz of z. Of particular interest for Hi-SEAS is the **return level** plot (and associated confidence intervals). The m^{th} return level is the value of the daily energy so *low* that it is *exceeded* once every m observations, here translated into numbers of years. We summary various results of special interest in the table below:

	Estimate	Lower confidence interval	Upper confidence interval
2 years	22.7kJ	17.9 kJ	27.9kJ
3 years	18.1kJ	12.2kJ	24.0kJ
4 years	15.2kJ	8.4kJ	22.0kJ
5 years	13.1kJ	5.6kJ	20.6kJ

As an example, once every 4 years, we expect to see a weekly maximum total daily flux of 8.4kJ or less.

Threshold excesses

Examining only the minima of the daily temperatures each week does not take full advantage of the dataset. For this reason, an alternative asymptotic analysis to extreme events is to determine the limiting distribution of $\mathcal{P}(X < u | X < u_0)$ where u_0 is a sufficiently large parameter and X represents the daily energy output. If the asymptotic limit of this distribution exists it takes the form of a generalised Pareto distribution,

$$H(y \equiv x - u) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \tag{4}$$

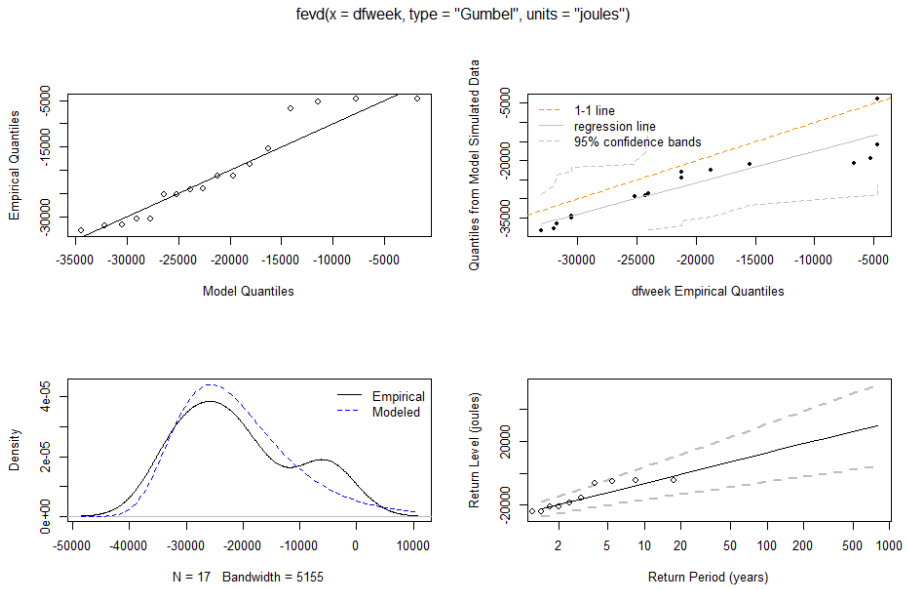


Figure 2: Diagnostics for the GEV fitting procedure.

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ and we fit the parameters ξ and $\tilde{\sigma}$ from the data. By empirically estimating $\zeta_{u_0} \equiv \mathcal{P}(X < u_0)$ one can assign probabilities to

$$\mathcal{P}(X > u > u_0) \approx \zeta_{u_0} H(x - u). \tag{5}$$

The threshold, u_0 must be sufficiently large that $\tilde{\sigma}$ is constant with respect to u. We found in our data that $u_0 \approx 35,000$ satisfies this condition within the 95% confidence interval. Our maximum likelihood analysis[2] yielded the following parameters

	μ	σ
Estimate	1.7×10^4	4.8×10^{-1}
Std Err	3.2×10^3	1.7×10^{-1}

It is an unfortunate artifact of the relatively small number of days in the dataset (120 days, approximately 17 weeks) and the rudimentary estimate of the total daily energy output that these parameters are not compatible with those of the GEV above.

Diagnostic plots provide the post-hoc justification of our analysis – figure (3). In particular, the probability distribution function provides estimates of finding $X - u$ within an infinitesimal region, dy of the value y. Return level plots give the value of the daily output that occurs once every m observations, here given in years. We summarise useful return levels in the following table:

	Estimate	Lower confidence interval	Upper confidence interval
2 years	19.4kJ	0kJ	56.9kJ
3 years	15.3kJ	0kJ	68.3kJ
4 years	12.8kJ	0kJ	77.4kJ
5 years	11.1kJ	0kJ	85.7kJ

For example, one expects a daily minimum in energy output of 15.3kJ once every three years.

[2] If K observations are above the threshold by amounts $y_i = x_i - u_0$ then the log likelihood is easily found to be

$$-K \log(\sigma) - (1 + \frac{1}{\xi}) \sum_{i=1}^{K} \log\left(1 + \frac{\xi y_i}{\tilde{\sigma}}\right) \tag{6}$$

with the $\xi = 0$ limit following from a limiting procedure.

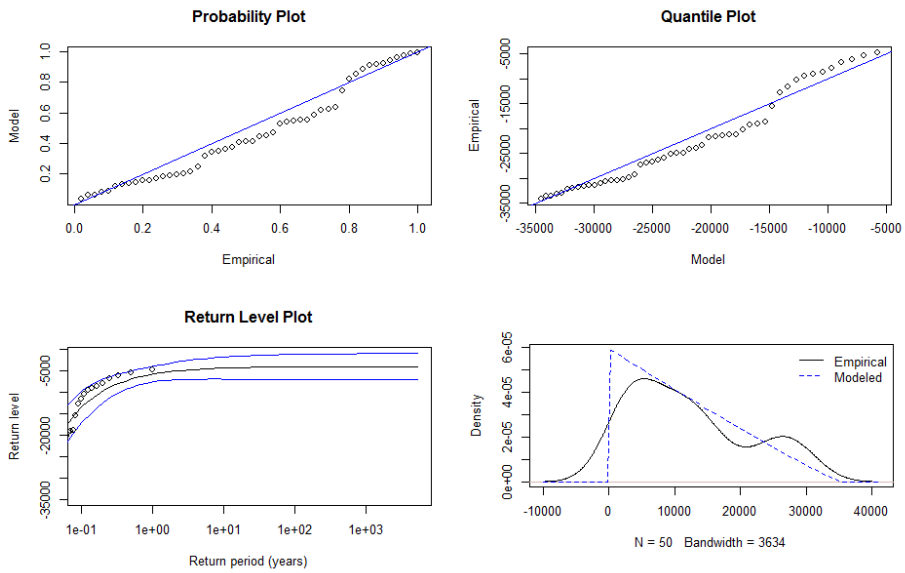


Figure 3: Diagnostics for Pareto fit.

Application and discussion

Our analysis provides a statistically rigorous means of anticipating extreme lows in solar panel energy output. We analysed the tails of the distributions describing the weekly minima in daily output as well as the asymptotic distribution of the data below a given threshold ($\approx$ 35kJ). Our diagnostics provide good post-hoc verification that the models are applicable and as such we take this study as "proof-in-principle" that this technique can be applied. Our results allow for planning of the backup or stored energy that is required to survive extreme events: we provide numerical estimates of how low daily energy output is likely to fall and how often one can expect this to happen (return levels), along with explicit probability distribution functions describing these extreme events. Incorporation of these probabilities should improve upon energy planning, energy storage and anticipation of extreme lows in energy output.

It is worth stating that the dataset is relatively small, once total daily energy output is calculated. However, the hourly measurements proved extremely useful in estimating this quantity. NASA provide data elsewhere that consists of day-averaged solar flux (amongst other things); whilst these data do go back over 20 years for some locations, the daily average flux will not provide as good an estimate of the daily energy generated, since the time integral is being replaced by a point estimate. Given more hourly data, covering many months or years, our analysis would be greatly improved[3] and we could provide more reliable estimates of the distributions and return levels.

Were we to have more time to analyse these data, we would wish to incorporate the additional information supplied by NASA. For example, one would expect that average daily temperate, say, would be a statistically significant factor in determining the total daily output of the radiation panels. More accurate estimates would incorporate these factors into our asymptotic models. Furthermore, simple linear regression would help to show up the dependence of the daily output on these factors, which could provide simple, non-asymptotic predictions of output for given weather conditions; these dependencies can then be used to refine our asymptotic models, by assigning simple regressive relationships of the scale, shape and location on the weather conditions, or by building hierarchical models that use these factors to produce estimates of the model parameters.

[3]The GEV and threshold models we have applied here require large numbers of data (for example, the GEV requires many blocks of data, each of which contains a large number of observations) for the analytic distributions to be realised by the observations.