# WCRP-ICTP Summer School on Attribution and Prediction of Extreme Events

## 21 July to 1 August 2014

Practice Sets for short course on R and the Extreme Value Analysis software package, `extRemes`

# 1 R Preliminaries

1. Create a matrix, assigned to `y`, with two columns of vectors containing 2, 1, 5 and 3, 7, 9, respectively. Give the matrix column names, and write it out to a file (**Hint**: See the help files for `colnames` and `write.table`).

2. Now write `y` out to a file as a comma separated file.

3. Read the files created in 1 and 2 above back into R, and assign them different names (**Hint**: See the help files for `read.table` and `read.csv`). See anything peculiar?

4. Check the class of these objects read from 3 above.

5. See the help file for the class type found in 4 above. This is a very important class in R. It is a cross between a matrix and a list object (see help files for these types as well). Unlike a list object, it must have the same numbers of rows for each column, and all columns must be a vector (e.g., a list can have wildly different component types, such as a function as one component, and a matrix as another). Unlike a matrix, a data frame can have different types of vectors across columns, such as a character vector in one column and a numeric vector in another.

6. Replace one of the entries for y with a missing value (i.e., NA).

7. Matrix multiply `y` by the vector `x` $< -$ `c( 1, 2, 0)`. That is, find $y^T x$ (**Hint**: See the help page for `%*%`, that is, type `?"%*%"` with the quotes, as well as for `t`). Now do the same for the `y` objects read in for 3 above. Anything unusual? Did you get an extra row for the csv file? If so, try writing it out again using `row.names=FALSE`. Matrix multiplication is not always possible with data frames. Convert the matrices read in to R to a matrix type object using the function `matrix`.

8. Install the package `fields`, if it is not already installed, and load this library into your R session.

9. Interpolate the daily maximum 8-hour ozone values for 18 June 1987 from the `fields` data set `ozone2` to a grid using thin-plate spline interpolation, and make a surface plot of the results (**Hint**: Run the examples in the help file for `fields`).

10. What class is the object `fit` from 7 above (i.e., `fit` is the name assigned in the help file for `fields`)?

11. See the help file for `surface`. Not very helpful, is it? See the help file for `surface.Krig` instead. Some functions, known as "method" functions, have specialized functions for different types of objects. Three very common examples are `predict`, `summary` and `plot`. List out all of the methods currently available for the function `plot` (**Hint**: See the help file for `methods`).

12. Methods are common when fitting a statistical model (e.g., a regression). List out all of the methods for objects of class "Krig." For objects of class "lm" ("lm" is the class associated with the main function in R for fitting linear models (e.g., linear regression), called by the same name, `lm`). **Hint**: be sure to specify the `class` argument here.

13. List out the objects in the current working directory for R (**Hint**: `ls()`).

14. Use `search()` to determine the position of the package `fields`, then use the `pos` argument of the `ls` command to list out the functions contained in the `fields` package.

15. Check the help file for formulas in R. Suppose I want to model a regression without an intercept, what do I need to add/subtract to the formula? Suppose I want to include sines and cosines?

## 2 Stationary Block Maxima

1. Simulate 500 maxima from samples of size 40 from the normal distribution.

2. Simulate 500 maxima from samples of size 40 from the exponential distribution.

3. Fit the GEV distribution to the simulated data from 1 above.

4. Fit the GEV distribution to the simulated data from 2 above.

5. Plot the QQ-Plot for the fit from 3 above. Do the assumptions for fitting the GEV distribution to these simulated data appear reasonable?

6. Plot the QQ-Plot for the fit from 4 above. Do the assumptions for fitting the GEV distribution to these simulated data appear reasonable?

7. In the lectures, all CI's reported used the normal approximation method. Estimate the 95% confidence intervals using the profile likelihood method for the shape parameter found from the fit in 3 above (**Hint**: use the arguments `which.par = 3` and `verbose = TRUE` in the call to `ci`. See also the help file for `ci.fevd`). Note that it is always a good idea to look at the profile likelihood plot (obtained using `verbose = TRUE` in the call to `ci`) to make sure the bounds are valid (the profile should cross the blue horizontal line at the same place as the vertical blue dashed lines). It may be necessary to use the `xrange` argument in order to get a more readable plot. The function will try to find a decent `xrange` by default, but will often fail, so good bounds may need to be found by the user using trial-and-error tactics. Is this parameter significantly different from zero at the 5% significance level?

8. Find the normal approximation interval for this same shape parameter. Do the bounds differ substantially from those of the profile likelihood method? Typically, the profile likelihood method is more appropriate for long return periods, and sometimes also for the shape parameter. The normal approximation method will always be symmetric and does not preserve the range of the parameter (e.g., one could acquire a lower bound for the scale parameter that is less than zero). The profile likelihood is range preserving, and can be asymmetric, which for longer return periods is usually more appropriate.

9. Simulate samples of size 10, 20, 50, 100 and 500 from the GEV distribution with location parameter 2, scale parameter 1.5 and shape parameter $-0.5$ (**Hint**: use `revd`).

10. Fit the GEV distribution to each sample from 9 above. Check the QQ-Plots for each fit, and estimate 95% CI's for the shape parameter in each case. Are the assumptions for the fits reasonable?

11. Load the `PORTw` data set from the package `extRemes`. What is the class of this data object? Use `colnames` or `names` to list the available fields.

12. See the help file for this data set to learn what each field represents.

13. Make a line plot of the maximum temperature against year for these data (**Hint**: use `type = "l"` in the plot call).

14. Make a scatter plot of the AO index against the maximum temperature. Does there appear to be much correlation?

15. Fit the GEV to the maximum temperature field.

16. Make a QQ-Plot for this fit. Do the assumptions for using the GEV appear reasonable for these data?

17. Estimate 95% CI's for the shape parameter. What can you say about the tail behavior of maximum temperature for Port Jervis, New York based on the fit to these data?

18. Make a line plot of the year against minimum temperature.

19. Fit the GEV to the minimum temperature data, and check the QQ-Plot (**Hint**: take the negative of the values using the syntax `-TMN0~1` as the first argument in the call to `fevd` (note that $\min\{x_1, \ldots, x_n\} = -\max\{-x_1, \ldots, -x_n\}$). Remember that the results are in terms of the maxima of negative values when you interpret them!

20. Estimate a 95% CI for the shape parameter. What can you say about minimum temperature for Port Jervis, New York based on these data?

21. Now analyze the minimum temperature series for Spet-Iles Québec, and answer the same questions as above for Port Jervis.

# 3 Stationary Peaks Over Threshold

1. Load the data set called `Fort`.

2. See the help file for this data set to learn what it contains.

3. What is the class for this data set?

4. Plot precipitation against `tobs`, `month`, `day` and `year`? Any patterns or trends?

5. Use `summary(Fort)` to see a summary of the data.

6. Fit the GP distribution to a range of thresholds, and select a threshold for fitting the GP distribution to these data. Does 0.395 inches appear to be a reasonable choice for a threshold?

7. Plot precipitation against `obs`, and add a red horizontal dashed line at 0.395 (**Hint**: use `abline` with argument `h = 0.395`). Do the data appear to be independent over the threshold?

8. Use `blockmaxxer` to obtain a similar data frame, but with annual maximum daily precipitation, and assign the new data frame to `Fort2`. The result should be a data frame with the annual maximum precipitation in the `"Prec"` column along with the position during each year where the maximum occurred.

9. Plot of these aggregated values against `year`, and add a red dashed horizontal line at 0.395 inches. Do the threshold excesses appear to be more independent now?

10. Fit a GP distribution to original Fort Collins, Colorado precipitation data, and fit the GEV distribution to the `Fort2` data set. Plot diagnostics for the resulting fits (i.e., use `plot`). How do the results compare?

11. Estimate a 95% CI for the shape parameter from both fits. Is the shape parameter significantly different from zero at the 5% level for either fit? What can you say about the results for the two fits? Which do you think is more reliable? Do you believe either one?

12. Estimate the extremal index using a threshold of 0.395 inches for the original precipitation field from `Fort`. Are the excesses independent? Decluster the excesses using the `decluster` function with a run length suggested from the call to `extremalindex`. Plot the results.

13. Re-fit the GP distribution to the newly de-clustered field. Is this fit any different from the previous ones? How does the shape parameter compare with the GEV distribution fit to the annual maxima, relative to the previous GP fit to the non-declustered excesses?

14. Estimate the Poisson rate parameter associated with a threshold of 0.395 inches for the (physically) de-clustered precipitation data (**Hint**: remember to use `> 0.395` in the call to `fpois` or `mean`).

15. Fit a PP model to the (physically) declustered precipitation data.

16. Find the Poisson rate parameter from the PP model fit above. Is it nearly the same as the estimate obtained above? **Hint**: use the relation $\hat{\lambda} = \left[1 + \frac{\hat{\xi}}{\hat{\sigma}}(u - \hat{\mu})\right]^{-1/\hat{\xi}}$).

17. Plot the diagnostics for the point process model fit. Do the assumptions for the model appear to be reasonable?

18. Try increasing the threshold to 2 inches. First, estimate the extremal index for the original `Fort` precipitation data. Are the excesses over this higher threshold independent, or reasonably so? If not, decluster the excesses using the suggested run length. If not, continue with the original data. Plot the diagnostics for the new fit. Now how do they look? Are all of the assumptions for the PP model reasonable using a threshold of 2 inches?

## 4 Linear temporal trends

1. Load the `Denmint` data set, and look at its help file.

2. Create a new data frame, called `Denmint2`, with a column added containing the negative of the minimum temperature, call it `negMin`, then use `blockmaxxer` to obtain the annual maxima of the negative of the minimum temperature, and assign this new data frame to `Denmint2`.

3. Plot `negMin` from `Denmint2` against year. Does there appear to be any temporal trend in these data?

4. Fit a linear regression of year against negative minimum temperature (**Hint**: See the help file for `lm`). Is there a significant linear trend in these data (**Hint**: use the `summary` function on the `lm` fitted object)?

5. Fit the GEV distribution to the annual maximum negative minimum temperature (without any trend).

6. Plot the diagnostics for this fit. Do the model assumptions appear to be reasonable?

7. Estimate a 95% CI for the shape parameter.

8. Interpret the return level plot for a gas/power company wanting to understand the risk of too much demand for gas in Denver in any given year (**Hint**: remember the return levels are for the negative of minimum temperature).

9. Fit the GEV to the negative minimum temperature data with a linear trend in the location parameter for $t = 1, 2, \dots$ (i.e., use the `Time` column).

10. Plot the diagnostics for this fit. Do the assumptions for the model fit appear to be reasonable?

11. Perform a likelihood ratio test for $\mu_1 = 0$ in the fit from 10 above. Is the result consistent with the result from the regression fit from 4 above?

12. Simulate 1000 maxima from normal distributions of size 30 with mean increasing at a rate of 25% (i.e., with slope 0.25), and standard deviation of 10. **Hint**: use

    ```
    z <- matrix(rnorm(1000 * 30,
        mean = rep(1:1000, each = 30) * 0.25, sd = 10), 1000, 30)
    ```

    and then apply the maximum to each row.

13. Plot of the resulting sample. Is there a trend?

14. Fit the GEV distribution to the sample without a trend.

15. Plot the diagnostics. Are the model assumptions reasonable here?

16. Fit the GEV to the sample with a linear trend in the location parameter (**Hint**: use `location.fun = ~I(1:1000)`).

17. Plot the diagnostics for the fit with a linear trend in the location parameter. Are the model assumptions reasonable?

18. Perform a likelihood ratio test for $\mu_1 = 0$ on this fit. Is inclusion of the linear trend statistically significant?

19. Try fitting the GEV to the sample with a linear trend in both the location and the scale parameters (using `use.phi = TRUE`), and check the diagnostic plots. Is this a reasonable model? Perform a likelihood-ratio test of this model against the previous fit with a trend. Is it statistically significant to include a trend in the scale parameter? What happens if the test is performed agains the model with no trends?

# 5 Cyclic variation

1. The GP distribution was fit to Fort Collins precipitation excesses, and declustered versions of these data, in the threshold excess practice above. Now, let's fit the Poisson rate parameter including an annual cycle using the `glm` function. This can be accomplished in the following way.

   ```
   Fort$PrecGTu <- Fort$Prec > 0.395
   fit <- glm(PrecGTu   sin(2 * pi * tobs / 365.25) +
       cos(2 * pi * tobs / 365.25), data = Fort, family = poisson())
   ```

   Is there a significant (at the 5% level) annual cycle in the Poisson rate parameter? Note that the model fit above is $\hat{\lambda}(t) = \hat{\lambda}_0 + \hat{\lambda}_1 \sin(2\pi t/24) + \hat{\lambda}_2 \cos(2\pi t/24)$, where $t = 1, \ldots, 365$.

2. Re-fit the PP model to the precipitation data with no parameter covariates and threshold of 0.395 inches, if you do not still have it.

3. Fit the PP model to the Fort Collins, Colorado precipitation data with a threshold of 0.395 mm, and with a cyclic variation in the location parameter as $\hat{\mu}(t) = \hat{\mu}_0 + \hat{\mu}_1 \sin(2\pi t/24)$ for $t = 1, \ldots, 365$.

4. Perform a likelihood ratio test for $\mu_1 = 0$ in the above model. Is the fit significant? Are the model assumptions reasonable?

5. Try fitting the point process model with a cyclic trend in the scale parameter (i.e., $\log \sigma(t) = \sigma_0 + \sigma_1 \sin(2\pi t/24) + \sigma_2 \cos(2\pi t/24)$). Is the trend significant? Are the model assumptions reasonable?

6. Given the results here, and the results from declustering previously, which approach would you recommend for these data?

# 6 More Practice

1. List out the arguments for the function, `optim` (**Hint**: use the `args` function).

2. See the help file for the function, `optim`.

3. Type `date` from the R prompt, and then hit return. What happens?

4. Now, type `date()` and hit return. What happens?

5. See the help file for `extRemes` to see, among other things, a list of the data sets included with the package.

6. Analyze the `Peak` data set. Is a block maxima or threshold excess model more appropriate here? Do there appear to be any tends in the data?

7. Analyze the maximum winter temperature for Sept-Iles. Do any of the other fields included with the data set make sense to try as covariates?