Supplementary Information for Appendix D

Richard Smith

The objective of this note is to specify the methods and programs used in the data analysis for the Appendix on Statistics of Extreme Events.

Data Downloading

Example 1: Precipitation patterns in southwest USA and Uruguay/Argentina as originally documented in the paper of Herweijer and Seager (2008).

Example 2: Monthly temperatures over Russia and precipitations over a region of Pakistan, motivated by the analysis of Lau and Kim (2012) – the same grid cells as they used were analyzed to construct a longer historical record.

The data source was the CRU gridded precipitation and temperature datasets available from

http://www.cru.uea.ac.uk/cru/data/hrg/cru_ts_2.10/

From this page, clicking on "data_all", the following files were downloaded:

cru_ts_2_10.1901-2002.pre.Z (137MB)

cru_ts_2_10.1901-2002.tmp.Z (118MB)

Data files represent monthly averages (1901-2002) of temperature and precipitation on a 720x360 grid – X values represent longitudes from -180 to +180, Y values represent latitudes from -90 to +90, units of 0.5 degrees in both. All values have to be multipled by 0.1 to convert to correct units. The four "boxes" used for this analysis are as follows:

Herweijer-Seager N American Box: 95-120W and 25-35N (Long. cells 121-170, lat. cells 231-250)
Herweijer-Seager S American Box: 50-65W and 30-40S (Long. cells 231-260, lat. cells 101-120)
Pakistan box: 70-73E, 32-35N (Long. cells 501-506, lat. cells 245-250)
Russia Box: 30-60E, 45-65N (Long. cells 421-480, lat. cells 271-310)

For each grid cell, a weight was calculated proportional to the cosine of latitude (to correct for varying area as latitude varies) and a weighted average computed over all cells within the specified box.

Herweijer-Seager data. The monthly values were further processed into annual means (simple average over 12 months for each year) and downloaded in the file "USA-Arg.txt". At this point of the analysis, the data were still in units of 10 mm/mon, so they still had to be multiplied by 0.1 to convert to correct units.

Code (in R) for Figure D-1 of report:

```
# Minima in USA-Arg dataset
z4=matrix(scan('USA-Arg.txt'),byrow=T,ncol=2)
z4=-0.1*z4
r2=rank(z4[,1])
r3=rank(z4[,2])
f2=-1/log((r2-0.5)/length(z4[,1]))
f3=-1/log((r3-0.5)/length(z4[,1]))
postscript('fig4.ps')
par(mfrow=c(1,2),pty='s',cex=1.5)
plot(-z4[,1],-z4[,2],xlab='SW-USA Rain (mm/mon)',ylab='Urug-Arg Rain (mm/mon)',pch=20)
text(25.5,56.0,'A')
plot(f2,f3,xlab='Transformed SW-USA Rain',ylab='Transformed Urug-Arg Rain',pch=20)
text(f2[10]+12,f3[10],'A')
text(f2[17]+12,f3[17],'A')
text(f2[24]+12,f3[24],'A')
dev.off()
```

Russia/Pakistan Data. June/July/August averages for each year were calculated and stored in the file "RusPak2.txt". The initial processing of the data (Figure D-2 of report) was as follows. The value for 2010 was not covered by the CRU dataset but was estimated independently from NCEP reanalysis; however this is only for illustration as the 2010 values are not used subsequently. The code differs slightly from the code for the USA-Argentina dataset as the focus is on maxima rather than minima.

```
z4=matrix(scan('RusPak2.txt'),byrow=T,ncol=2)
r2=rank(z4[,1])
r3=rank(z4[,2])
f2=-1/log((r2-0.5)/length(z4[,1]))
f3=-1/log((r3-0.5)/length(z4[,1]))
plot(f2,f3)
postscript('fig3.ps')
par(mfrow=c(1,2),pty='s',cex=1.5)
plot(z4[,1],z4[,2],ylab='Pakistan Rain (mm/mon)',xlab='Russian Temperature (deg C)',
xlim=c(min(z4[,1]),21.8),pch=20)
points(21.73,126.7,pch='*',col='red',cex=1.5)
text(21.15,126.7,'2010',col='red')
plot(f2,f3,ylab='Transformed Pakistan Rain',xlab='Transformed Russian Temperature',pch=20)
dev.off()
```

Logistic Model Analysis

The analysis was based on data transformed to unit Fréchet by a rank procedure (same as for the two figures) and likelihood function computed by portioning the sample space into four quadrants. Here f2, f3 are the Fréchet-transformed data and u2, u3 the thresholds:

```
lh1=function(b){
if(b<=0|b>=1){
lh1=10^10
}
if(b>0&b<1){
lh1=0
for(i in 1:length(f2)){
if(f2[i]<=u2&f3[i]<=u3){
lh1=lh1+(u2^(-1/b)+u3^(-1/b))^b
}
if(f2[i]>u2&f3[i]<=u3){
lh1=lh1+
(1/b+1)*log(f2[i])-(b-1)*log(f2[i]^(-1/b)+u3^(-1/b))+(f2[i]^(-1/b)+u3^(-1/b))^b
}
if(f2[i]<=u2&f3[i]>u3){
lh1=lh1+
(1/b+1)*log(f3[i])-(b-1)*log(u2^(-1/b)+f3[i]^(-1/b))+(u2^(-1/b)+f3[i]^(-1/b))^b
}
if(f2[i]>u2&f3[i]>u3){
lh1=lh1-
log((f2[i]^(-1/b)+f3[i]^(-1/b))^b+1/b-1)+
(1/b+1)*log(f2[i]*f3[i])-
(b-2)*log(f2[i]^(-1/b)+f3[i]^(-1/b))+
(f2[i]^(-1/b)+f3[i]^(-1/b))^b
}
}
}
lh1
}
```

There is a separate function to compute the ratio of joint to independent probabilities at different probability levels:

```
pr1=function(b,N){
z1=-1/log(1-1/N)
z2=-1/log(1-1/N)
# independent case
p1=exp(-1/z1)
p2=exp(-1/z2)
p3=p1*p2
```

```
p4=1-p1-p2+p3
# dependent case
p1=exp(-1/z1)
p2=exp(-1/z2)
p3=exp(-(z1^(-1/b)+z2^(-1/b))^b)
p5=1-p1-p2+p3
p5/p4
}
```

The method was to specify a single threshold for both variables (=2.5 on Fréchet scale) and optimize using the optim function in R:

```
thr=2.5
u2=thr
u3=thr
opt1=optim(0.5,lh1,method='BFGS')
opt1$par
pr1(opt1$par,10)
pr1(opt1$par,20)
pr1(opt1$par,50)
```

Results for USA-Argentina dataset:

```
> opt1$par
[1] 0.8469012
> pr1(opt1$par,10)
[1] 2.736862
> pr1(opt1$par,20)
[1] 4.748002
> pr1(opt1$par,50)
[1] 10.78752
```

Results for Russia-Pakistan dataset:

```
> opt1$par
[1] 0.9992513
> pr1(opt1$par,10)
[1] 1.008856
> pr1(opt1$par,20)
[1] 1.019215
> pr1(opt1$par,50)
[1] 1.050334
```

Next, we created confidence intervals via a bootstrap. The original data was resampled with replacement, the Fréchet transform recalculated, and the logistic model re-estimated. Finally the estimated probability ratios were ordered and the 5th and 95th bootstrap quantiles were calculated.

```
# results of bootstrap analysis for USA-Argentina data
> boot2[,2:4]
         [,1]     [,2]      [,3]
[1,] 1.287815 1.623912  2.633541
[2,] 4.097263 7.655521 18.338296

# results of bootstrap analysis for Russia-Pakistan data
> boot2[,2:4]
         [,1]     [,2]     [,3]
[1,] 1.000893 1.001937 1.005075
[2,] 1.011730 1.025450 1.066668
```

Ramos-Ledford model

The likelihood function was coded as in Ramos and Ledford (2009):

```
lh3=function(pars){
eta=pars[1]
rho=pars[2]
alf=pars[3]
lam=pars[4]
if(rho<=0|alf<=0|eta<=0|eta>1|lam<=1-2*exp(-1/thr)|lam>=1){
lh3=10^10
}
if(rho>0&alf>0&eta>0&eta<=1&lam>1-2*exp(-1/thr)&lam<1){
lh3=0
Nrho=rho^(-1/eta)+rho^(1/eta)-(rho^(-1/alf)+rho^(1/alf))^(alf/eta)
for(i in 1:length(f2)){
if(f2[i]<=thr&f3[i]<=thr){
den=2*exp(-1/thr)-1+lam
if(is.na(den)==F&den>0){lh3=lh3-log(den)}
if(is.na(den)==T|den<=0){lh3=10^10}
}
if(f2[i]>thr&f3[i]<=thr){
den=exp(-1/f2[i])/f2[i]^2-lam*thr^(1/eta)*rho*(
(rho*f2[i])^(-1-1/eta)-((rho*f2[i])^(-1/alf)+(thr/rho)^(-1/alf))^(alf/eta-1)*
(rho*f2[i])^(-1-1/alf))/(eta*Nrho)
if(is.na(den)==F&den>0){lh3=lh3-log(den)}
if(is.na(den)==T|den<=0){lh3=10^10}
}
if(f2[i]<=thr&f3[i]>thr){
den=exp(-1/f3[i])/f3[i]^2-lam*thr^(1/eta)*(
(f3[i]/rho)^(-1-1/eta)-((f3[i]/rho)^(-1/alf)+(rho*thr)^(-1/alf))^(alf/eta-1)*
(f3[i]/rho)^(-1-1/alf))/(eta*rho*Nrho)
if(is.na(den)==F&den>0){lh3=lh3-log(den)}
```

```
if(is.na(den)==T|den<=0){lh3=10^10}
}
if(f2[i]>thr&f3[i]>thr){
den=lam*thr^(1/eta)*(eta-alf)*((rho*f2[i])^(-1/alf)+(f3[i]/rho)^(-1/alf))^(alf/eta-2)*
(f2[i]*f3[i])^(-1-1/alf)/(alf*eta^2*Nrho)
if(is.na(den)==F&den>0){lh3=lh3-log(den)}
if(is.na(den)==T|den<=0){lh3=10^10}
}
}
}
lh3
}
```

Again, the threshold was fixed at 2.5. The numerical difficulties for optimizing this likelihood are considerably greater than those for the logistic likelihood. As a test, the optimization algorithm was run six different ways (3 starting values and 2 methods). The results were identical for the USA-Argentina data and similar but not identical for the Russia-Pakistan data:

```
# test optimization with different starting values and optimization methods
# to ensure stability
opt1=optim(c(0.5,1,0.7,0.01),lh3,method='Nelder-Mead')
opt2=optim(c(0.5,1,0.7,0.01),lh3,method='BFGS',control=list(maxit=10000,ndeps=rep(10^(-6),4)))
opt3=optim(c(0.5,1,1,0.01),lh3,method='Nelder-Mead')
opt4=optim(c(0.5,1,1,0.01),lh3,method='BFGS',control=list(maxit=10000,ndeps=rep(10^(-6),4)))
opt5=optim(c(0.5,1,0.3,0.01),lh3,method='Nelder-Mead')
opt6=optim(c(0.5,1,0.3,0.01),lh3,method='BFGS',control=list(maxit=10000,ndeps=rep(10^(-6),4)))
print(c(opt1$value,opt2$value,opt3$value,opt4$value,opt5$value,opt6$value))
# function to calculate ratio of probabilities from Ramos-Ledford model
pr2=function(pars,N){
z1=-1/log(1-1/N)
z2=-1/log(1-1/N)
# independent case
p1=exp(-1/z1)
p2=exp(-1/z2)
p3=p1*p2
p4=1-p1-p2+p3
# dependent case
p1=exp(-1/z1)
p2=exp(-1/z2)
eta=pars[1]
rho=pars[2]
alf=pars[3]
lam=pars[4]
Nrho=rho^(-1/eta)+rho^(1/eta)-(rho^(-1/alf)+rho^(1/alf))^(alf/eta)
s=z1/thr
```

```
t=z2/thr
p5=lam*((rho*s)^(-1/eta)+(t/rho)^(-1/eta)-((rho*s)^(-1/alf)+(t/rho)^(-1/alf))^(alf/eta))/Nrho
p5/p4
}
print(c(pr2(opt1$par,10),pr2(opt2$par,10),pr2(opt3$par,10),pr2(opt4$par,10),pr2(opt5$par,10),pr2(opt6$par,10)))
print(c(pr2(opt1$par,20),pr2(opt2$par,20),pr2(opt3$par,20),pr2(opt4$par,20),pr2(opt5$par,20),pr2(opt6$par,20)))
print(c(pr2(opt1$par,50),pr2(opt2$par,50),pr2(opt3$par,50),pr2(opt4$par,50),pr2(opt5$par,50),pr2(opt6$par,50)))

# results with USA-Argentina dataset
> print(c(opt1$value,opt2$value,opt3$value,opt4$value,opt5$value,opt6$value))
[1] 335.2007 335.2007 335.2007 335.2007 335.2007 335.2007
> pr2(opt2$par,10)
[1] 2.885823
> pr2(opt2$par,20)
[1] 4.852135
> pr2(opt2$par,50)
[1] 9.876563

# results with Russia-Pakistan dataset
> print(c(opt1$value,opt2$value,opt3$value,opt4$value,opt5$value,opt6$value))
[1] 336.6401 336.6401 336.6612 336.7617 336.6401 336.8340 >
>print(c(pr2(opt1$par,10),pr2(opt2$par,10),pr2(opt3$par,10),pr2(opt4$par,10),pr2(opt5$par,10),pr2(opt6$par,10)))
[1] 0.3909917 0.3903157 0.4330005 0.3334259 0.3904195 0.3139739
>print(c(pr2(opt1$par,20),pr2(opt2$par,20),pr2(opt3$par,20),pr2(opt4$par,20),pr2(opt5$par,20),pr2(opt6$par,20)))
[1] 0.2665158 0.2658511 0.3152241 0.2086712 0.2659673 0.1903669
>print(c(pr2(opt1$par,50),pr2(opt2$par,50),pr2(opt3$par,50),pr2(opt4$par,50),pr2(opt5$par,50),pr2(opt6$par,50)))
[1] 0.1685981 0.1680108 0.2171388 0.1181898 0.1681215 0.1034883
# note: in this case the values from the six optimizations are not all the same,
# but opt1, opt2 and opt5 are all the same and have been used for the table in the report
```

We also ran a bootstrap procedure to get confidence intervals (this took about 2 hours to run with 10,000 bootstrap samples – could reduce that number and still get very similar results):

```
nsim=10000
i1=c(500,9501)
boot1=matrix(NA,ncol=7,nrow=nsim)
n=length(f2)
for(isim in 1:nsim){
ind1=rep(NA,n)
```

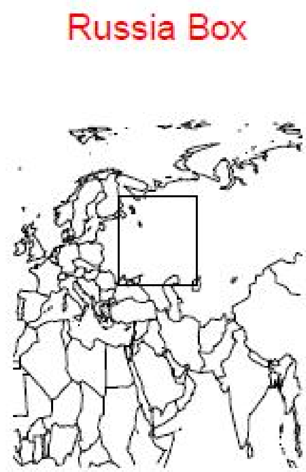
```
for(i in 1:n){ind1[i]=sample(1:n,1)}
r2a=rank(z4[ind1,1])
r3a=rank(z4[ind1,2])
f2=-1/log((r2a-0.5)/length(z4[,1]))
f3=-1/log((r3a-0.5)/length(z4[,1]))
#opt1=optim(c(0.5,1,1,0.01),lh3,method='Nelder-Mead')
opt1=optim(c(0.5,1,1,0.01),lh3,method='BFGS',control=list(maxit=10000,ndeps=rep(10^(-6),4)))
boot1[isim,1:4]=opt1$par
boot1[isim,5]=pr2(opt1$par,10)
boot1[isim,6]=pr2(opt1$par,20)
boot1[isim,7]=pr2(opt1$par,50)
}
boot2=matrix(NA,ncol=7,nrow=2)
for(j in 1:7){boot2[,j]=sort(boot1[,j])[i1]}
boot2[,5:7]

# results for USA-Argentina (10000 bootstrap simulations)
> boot2[,5:7]
         [,1]     [,2]      [,3]
[1,] 1.163495 1.233728  1.375366
[2,] 4.980126 9.584455 23.414172
# results for Russia-Pakistan (10000 bootstrap simulations)
boot2[,5:7]
           [,1]        [,2]        [,3]
[1,] 0.03770607 0.008292657 0.001214635
[2,] 1.35277240 1.837110427 2.920528207
```

Rough maps of the regions covered by these analyses:


American Southwest


Argentina/Uruguay


Pakistan Box


Russia Box