Non–stationary modelling of extremes of precipitation and temperature over mountainous areas under climate change

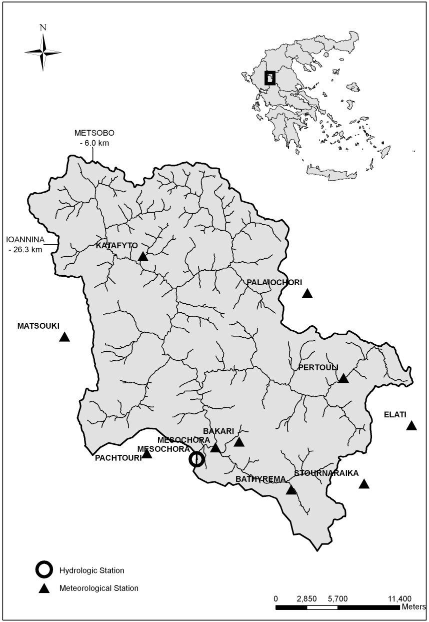
C. Caroni[1], D. Panagoulia[2] and P. Economou[3]

[1] Department of Mathematics, National Technical University of Athens,
[2] School of Civil Engineering, National Technical University of Athens,
[3] Department of Civil Engineering, University of Patras

ICRA 6, Barcelona, May 26–29, 2015

tu-logo

ur-logo


N
METSOBO
- 6.0 km
IOANNINA
- 26.3 km
KATAFYTO
PALAIOCHORI
MATSOUKI
PERTOULI
ELATI
BAKARI
MESOCHORA
MESOCHORA
PACHTOURI
BATHYREMA
STOURNARAIKA
Hydrologic Station
Meteorological Station
0
2,850
5,700
11,400
Meters

- There is a great deal of scientific literature in many fields on modelling extremes in time series.
- For example, in meteorology: modelling extremes of rainfall, temperature, wind speed. This is often related to engineering design: what is the greatest stress that a structure must withstand?
- For many years, since Jenkinson (*Quart. J. Roy. Met. Soc.*, 1955), the Generalised Extreme Value distribution has been employed to model meteorological extremes.

tu-logo

ur-logo

The GEV distribution

$$F(x;\mu,\sigma,\xi) = \begin{cases} \exp\left\{-\left[1+\frac{\xi(x-\mu)}{\sigma}\right]^{-1/\xi}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\frac{(x-\mu)}{\sigma}\right]\right\}, & \xi = 0 \end{cases}$$

with location $\mu \in \mathbb{R}$; scale $\sigma > 0$; shape $\xi \in \mathbb{R}$

- $\xi = 0 \;\Rightarrow$ Gumbel distribution with support $x \geq \mu + \frac{\sigma}{\xi}$
- $\xi > 0 \;\Rightarrow$ Fréchet distribution with support
- $\xi < 0 \;\Rightarrow$ (reversed) Weibull distribution with support $x \leq \mu - \frac{\sigma}{\xi}$

u-logo

logo

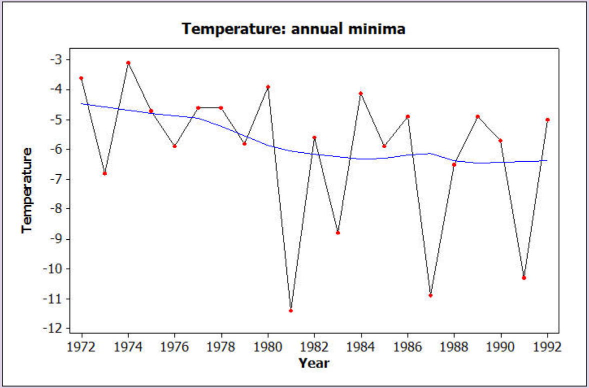
Non–stationarity

A series of maximum temperatures or maximum precipitation in an area might not be stationary, but could show trends over a period of time.

Indeed, there is mounting evidence that hydro–climatic extreme series are not stationary, owing to natural climate variability or anthropogenic climate change

Jain and Lall, *Water Resources Research*, 2001
Milly et al., *Science*, 2008.

Annual minimum temperatures in a mountainous area of Greece (with trend fitted by lowess in Minitab)



Modelling **non-stationarity** within the framework of the GEV distribution requires extended models with covariate-dependent changes in at least one of the distribution's three parameters (Coles, "*Intro. to Statist. Modelling of Extreme Values*", 2001).

Non–stationary GEV

Parameters are a function of time t and possibly other covariates as well (Coles, 2001). In the environmetrics literature, it is common to keep the shape ξ constant. In this case, the non-stationary over time GEV distribution is

$$F(y;\mu(t),\sigma(t),\xi) = \exp\left\{-\left[1+\xi\,\frac{y-\mu(t)}{\sigma(t)}\right]^{-1/\xi}\right\}$$

Nogaj et al., *Nonlin Proc Geophys*, 2007
El Adlouni et al., *Water Resources Research*, 2007
Cannon, *Hydrol. Proc.*, 2010

For example, the following polynomial regression structures could be considered for location and scale parameters when time is the explanatory covariate

$$\begin{aligned} \mu(t) &= \mu_0 + \mu_1 t + \mu_2 t^2 + \mu_3 t^3 \\ \sigma(t) &= \exp(\sigma_0 + \sigma_1 t + \sigma_2 t^2 + \sigma_3 t^3) \end{aligned}$$

allowing up to cubic dependence on time of both the location μ and scale σ parameters.

Denote by GEV*jk* the model with time dependence of order j in the location parameter and order k in the scale parameter. e.g. the GEV*21* non–stationary model assumes

$\hookrightarrow$ a quadratic trend ($\mu_3 = 0$) in location and

$\hookrightarrow$ a log–linear trend in scale ($\sigma_2 = \sigma_3 = 0$).

The stationary GEV distribution is GEV*00* .

tu-logo

ur-logo

Generalised Additive Models for Location, Scale and Shape (GAMLSS)

Rigby and Stasinopoulos, 2005, Applied Statistics

This class of models allows covariate–dependence in up to four parameters of a distribution chosen from a very wide family.

For example, Villarini et al., (*Adv. Water Resources*, 2010) examined the fit of five distributions (Gumbel, Weibull, Gamma, Logistic and Lognormal) to data on rainfall and temperature in Rome.

Non-stationary GEV distributions can also be fitted within this framework.

General model for parameter θ_k is

$$g_k(\theta_k) = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk}$$

where g_k is a link function, X_k is a design matrix containing the values of J_k covariates for each of n independent observations, β_k is a parameter vector of length J_k, Z_{jk} is another known design matrix of dimension $n \times q_{jk}$ and γ_{jk} is a q_{jk} - dimensional random vector.

GAMLSS modelling is implemented in the R package “gamlss” (http://cran.R-project.org/package=gamlss), which makes it easy to include features such as random effects or non–polynomial dependence on covariates by means of splines.

Model selection

- Given the availability of many alternative models, we need objective procedures for selecting which of various candidate models fits best (e.g. non–stationary instead of stationary? non–stationary in which parameters?)
- The likelihood ratio test can be used to compare two hierarchically nested models

tu-logo

ur-logo

Model selection

- Information criteria: choose model with smallest value of

$$AIC_C = -2\hat{\ell} + 2p + \frac{2p(p+1)}{n-p-1}$$

  (this is the corrected AIC - the third term is a small-sample adjustment) or

- $$BIC = -2\hat{\ell} + p \ln n$$

  where $\hat{\ell}$ is the maximized value of the likelihood from a model that contains p parameters, and n is the sample size

Empirical comparison of AICc and BIC for selection of GEV models

Simulation study to find how often each criterion correctly identifies the true model among the set of models GEVjk ($j = 0, 1, 2, 3$; $k = 0, 1, 2, 3$), for samples of sizes $n = 20, 50$ or 100 (Panagoulia et al., *Environmetrics*, 2014).

- True models: GEV*00*, GEV*10*, GEV*01*, GEV*11*, each with ξ = -0.1, 0 and 0.1
- Coefficients in non-stationary models equal to one
- 1000 generated samples for each combination of true model and sample size
- Fit all the GEVjk models to each sample, $j = 0, 1, 2, 3$, $k = 0, 1, 2, 3$ by maximum likelihood using R package "ismev"
- Select best according to AICc and BIC

Simulation results

- Results did not depend on shape ξ
- Both criteria had high success rates in detecting non–stationarity.
- BIC was more successful in identifying the correct model:
  - $\hookrightarrow$ $>$ 80% of the time for $n = 50$
  - $\hookrightarrow$ $>$ 90% for $n = 100$.
- Neither performed very well for $n = 20$, which is a very small sample in relation to the number of parameters in some of these models.

Confidence intervals

- Confidence intervals are often an important output from the fitting of models to series of extremes.
- Especially, we require confidence intervals for extreme quantiles of the distribution, which are usually expressed as return periods.

tu-logo

ur-logo

Return periods

- For example, annual rainfall data, event = year's rainfall exceeds the upper $q\%$ point

  $\Rightarrow$ P(event)=q in any given year

  $\Rightarrow$ E(years until event) = $1/q$ (geometric distribution)
  e.g. if $q = 0.01$, then,

  E(years until event) = 100, a 100-year return period.

- Obviously, estimates of these extreme quantiles can only be obtained by model fitting, because by definition they correspond to rare events.

Confidence intervals for a quantile

How to construct a confidence interval for a quantile? Any quantile = f(max. likelihood estimates of model parameters)

- Base CI for quantile on asymptotic normal distribution of MLEs. Unlikely to work for extreme quantiles
- Bootstrap methods
- Kysely (*J. Appl. Met. Clim.*, 2008) looked at this for stationary GEV
- Panagoulia et al. (*Environmetrics*, 2014) for non–stationary GEV

Simulation study

We compared in our simulation study 12 confidence intervals for quantiles.

These were obtained from all the combinations of

- 3 ways of constructing bootstrap samples
- 4 methods for constructing confidence intervals based on the bootstrap samples

Bootstrap samples

Parametric resampling

- Fit the model to the actual data
- Generate samples from this model, with parameter values equal to the estimates

Random-t resampling (case resampling)

- Construct each bootstrap sample by simple random sampling with replacement from the original data
- Entire data vectors (cases) are sampled

Fixed–t resampling

- The covariate vectors in the bootstrap samples are the same as in the original data
- The residuals from the fit to the original data are resampled
- The original covariate vector and the resampled residual give a generated value for each case

Constructing confidence interval

- Normal - the CI is constructed using Normal distribution with mean, sd obtained from bootstrap distribution
- Percentile - non-parametric CI, from the order statistics of the bootstrap distribution
- Basic - transformation of percentile interval
- Bias-corrected and accelerated (BCa) - Modify percentile CI to correct for bias and skewness

Illustrative results (1):

Simulated coverage proportion of various bootstrap 95% confidence intervals of estimate of upper 1% point of GEV*00* ($\mu = 0$, $\sigma = 1$, $\xi = -0.1$). 1000 runs, simulated sample size *n*=100 and 1000 bootstrap samples in each run.

	Normal	Basic	Percentile	BCa
Fixed - *t*	0.896	0.884	0.875	0.910
Random - *t*	0.891	0.885	0.877	0.912
Parametric	0.910	0.899	0.881	0.918

BCa does best

Summary of results for simulated CIs

- No clear differences between methods concerning CIs for parameters
- Intervals very wide for extreme quantiles
- Coverage probabilities well below nominal level for quantiles from the 90th onwards - except for BCa method
- Parametric method a bit better than random-t and fixed-t, which hardly differ
- Computation time similar for all the methods

Illustrative results (2):

Simulated coverage proportion of BCa parametric bootstrap 95% confidence intervals estimate of upper 1% point for various GEV models, with $n = 50$.

	GEV*00*	GEV*10*	GEV*01*	GEV*11*
ξ = -0.1	0.914	0.927	0.935	0.976
ξ = 0	0.913	0.932	0.933	0.969
ξ = 0.1	0.912	0.931	0.915	0.964

Values in range 0.936-0.964 are not significantly different from nominal 0.95

tu-logo

logo

Illustrative results (3):

Simulated coverage proportion of BCa parametric bootstrap 95% confidence intervals estimate of upper 1% point for various GEV models, with $\xi = -0.1$.

	GEV*00*	GEV*10*	GEV*01*	GEV*11*
$n = 20$	0.971	0.966	0.979	0.990
$n = 50$	0.914	0.927	0.935	0.976
$n = 100$	0.918	0.932	0.930	0.963

Values in range 0.936-0.964 are not significantly different from nominal 0.95

tu-logo

logo

Application

- Data from a river catchment in a mountainous area of Greece
- Historical data 1972-1992
- Data under two climate-change scenarios, 1961-2000 and 2061-2100
- Annual maximum precipitation over whole area was analysed in Panagoulia et al. (2014).
- Stationary GEV model for precipitation was supported. Furthermore, $\xi = 0$ not rejected (i.e. Gumbel distribution)

Here we look at temperature

u-logo

logo

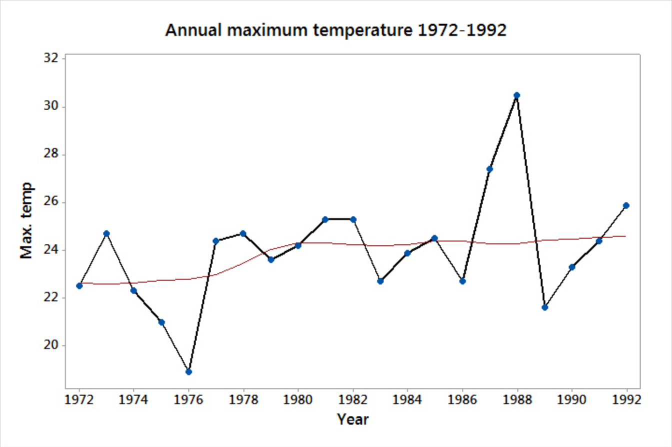
Application - annual maxima - minima of temperature

- Analysis presented here: annual maxima and minima of historical temperature data
- Overall and separately in 9 zones (dividing the area by elevation)
- Model dependence on time and space
- Use other distributions from GAMLSS as well as the GEV

tu-logo

ur-logo

Results: annual maxima of temperature



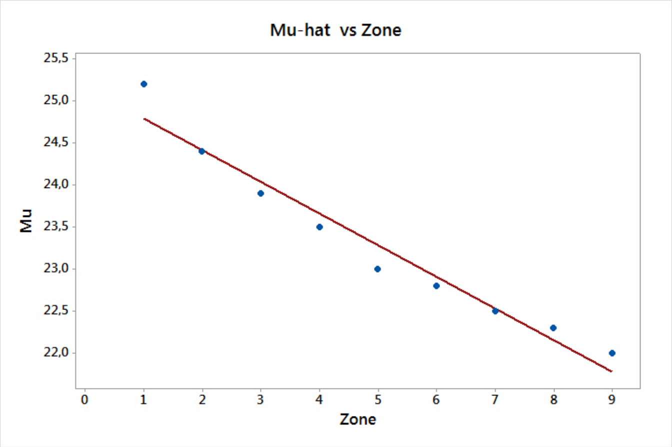
Results: annual maxima of temperature

MLE of stationary GEV separately in each zone

Zone (altitude)	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	25.2	2.05	-0.23
2	24.4	2.05	-0.19
...	...	...	...
5	23.0	2.22	-0.17
...	...	...	...
9	22.0	3.06	0.06

- ↪ Strong suggestion of spatial dependence
- ↪ Also, some evidence ($p = 0.05$ in likelihood ratio test) supporting GEV*10* model in Zone 5 and higher, i.e. μ depending linearly on time

Plot of $\hat{\mu}$ versus elevation (zone)



Fit GEV to data on all zones, introducing zone as a covariate for the parameters - annual maxima of temperature

Preferred model (AICc, BIC, LR tests)

- μ depends on zone and year
- $\ln \sigma$ depends on zone
- ξ constant over zones and years

$$\hat{\mu} = \underset{(0.36)}{25.00} - \underset{(0.073)}{0.285}\, Zone + \underset{(0.030)}{0.118}\,(Year - 1982)$$

$$\ln \hat{\sigma} = \underset{(0.109)}{0.554} + \underset{(0.021)}{0.068}\, Zone$$

$$\hat{\xi} = \underset{(0.045)}{-0.142}$$

- sig. different from zero (unlike models for rainfall)

Similar analysis for annual minima of temperature

Carry out analysis as for max, with input = (-1) x min

- Preferred model: μ depends on zone and year
- $\ln \sigma$ depends on year
- ξ constant over zones and years

$$\hat{\mu} = \underset{(0.314)}{0.438} + \underset{(0.060)}{1.043}\,\mathit{Zone} + \underset{(0.024)}{0.121}\,(\mathit{Year} - 1982)$$

$$\ln \hat{\sigma} = \underset{(0.063)}{0.579} + \underset{(0.011)}{0.026}\,(\mathit{Year} - 1982)$$

$$\hat{\xi} = \underset{(0.063)}{0.004} \quad \text{- not sig. different from zero}$$

Other distributions

- Use R package "gamlss" to try fitting other distributions to these data
- In particular, Gamma, Log Normal, Inverse Gaussian
- These are 2 parameter distributions: allow both parameters to depend on covariates
- We show results of fitting to maxima annual temperatures including zone effects

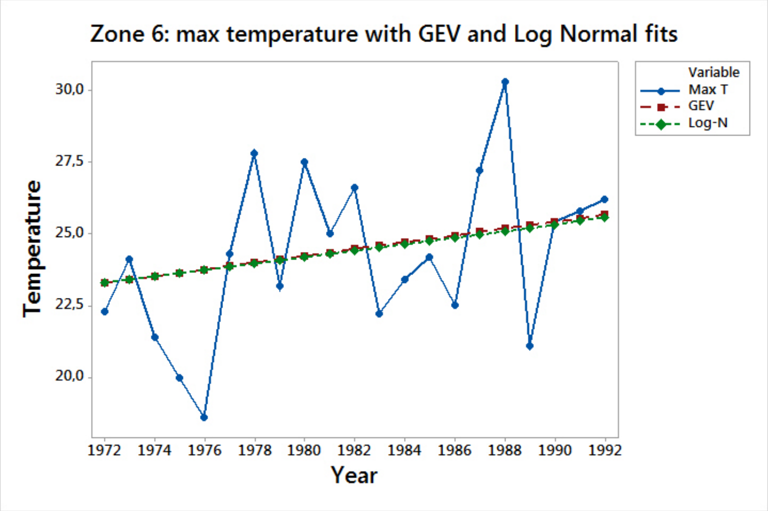
u-logo

ur-logo

- There seems to be very little difference between the overall fits of these distributions
- For the preferred model (one parameter depending on year and zone, the other depending on zone - same model as GEV), BIC values are

Gamma	928.25
Inverse Gaussian	927.38
Log Normal	926.77

- The following slide shows the closeness of fits between GEV and Log Normal


Zone 6: max temperature with GEV and Log Normal fits
Temperature
30,0
27,5
25,0
22,5
20,0
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
Year
Variable
Max T
GEV
Log-N

Conclusions (1)

- Fitting GEV to stationary series of extremes is a well-established methodology in the analysis of climate data and elsewhere.
- The evidence shows that in fact series are often not stationary.
- Non-stationarity can be catered for by allowing the distribution's parameters to depend on time and other covariates.
- Model selection is important: we carried out a study of the use of information criteria to select the best model. We found that the BIC works best, except for small n when AIC is preferable.
- These criteria detect non-stationarity with high success.

Conclusions (2)

- Often, the quantities of main interest are quantiles estimated from the fitted model, especially extreme quantiles corresponding to long return periods.
- We require confidence intervals for these estimates. We investigated the accuracy of bootstrap CI's, for 12 combinations of methods of construction and ways of drawing bootstrap samples.
- We recommend BCa CI's as much better than CI constructed by the other methods, and parametric resampling for bootstrapping as slightly better than other techniques.

u-logo

ur-logo

Conclusions (3)

- Other distributions besides GEV can be used
- The GAMLSS framework makes possible the fitting of other distributions in non-stationary form, with up to four parameters depending on covariates
- In our application, GEV and GAMLSS modelling led to the same conclusions concerning dependence on time and space

tu-logo

ur-logo

Conclusions (4)

- Different distributions (e.g. GEV and Log Normal) produced almost indistiguishable fits in this example.
- Perhaps surprisingly, estimates of extreme quantiles were also similar
  e.g. 99.9^{th} percentile (Zone 5, 1982)

GEV	34.3
Gamma	33.6
Inverse Gaussian	34.1
Log Normal	34.1

u-logo
ur-logo