Statistical downscaling of precipitation using extreme value theory

Petra Friederichs

Meteorological Institute
University of Bonn

E2C2-GIACS Summer School, Romania, September 3-12, 2007
Outline

Statistical downscaling – general overview
- Aim of statistical downscaling
- Approaches and Methods
- with regard to extremes

Our Problem
- Data
- Probabilistic Downscaling
- Extreme value theory
- Censored quantile regression
- Verification

Results
- Extreme precipitation events
- Skill of downscaling

Conclusions
Weather Forecasting

**DYNAMICS:**
In numerical weather forecasting the **Navier Stokes equations** on a rotation sphere and the $1^{\text{st}}$ principle of thermodynamics are numerically integrated in time.

**PHYSICS:**
- sub-scale processes are parametrized
  - cloud micro-physics
  - radiative transfer
  - exchange processes with land surface...
  - turbulence
  - convection ...

Extreme weather phenomena are (mostly) on small scales

$\rightarrow$ **post-processing**

from DWD
Aim of downscaling:

Represent and forecast local weather or climate conditions on scales not resolved by the model. Here: downscaling in space!

**Dynamical** and **statistical** downscaling

Dynamical downscaling uses a **regional, high resolution model**, where the boundary conditions are determined by the global model.

Statistical downscaling derives a **statistical model** between the large scale model state and the local conditions.
Downscaling

Applications:

**Short range weather prediction**
- regional models to account for local conditions (orography) and small scale processes (convection)
- statistical downscaling (MOS) for local (weather station) forecasts
- assessing probabilities e.g. of extreme events

**Climate change projections**
- estimate possible man induced changes on the local scale

**Weather generator**
- generate realistic precipitation patterns for hydrological modeling
Methods of Stat. Downscaling

Linear transfer functions
- multiple linear regression
- canonical correlation analysis

Non-linear transfer functions
- neural networks

Stochastic models
- multivariate autoregressive models
- conditional weather generator

Weather typing methods
- conditional resampling
- analogue-based methods

Probabilistic downscaling
- estimate conditional distribution function
Extreme Weather Phenomena

**heavy precipitation events:**
reasonable observational network
spacial dependence – unknown, depends on weather

2001 May 03
convective cell

2005 Jan 02
frontal precipitation

2004 Jul 17
frontal and stratiform

**wind gusts:**
must less observations
processes not well known
spacial dependence - unknown

**heat waves:**
good observations
but long memory
Data

Period from: 1948 to 2004
Separate: cold (NDJFM) and warm seasons (MJJAS)
About: 8600 daily totals

NCEP reanalysis
Daily mean fields of $\omega_{850}$, $\zeta_{850}$ and PWat

DWD (German Weather Service)
~ 2000 stations in Germany
Daily precipitation totals
Downscaling

Downscaling seeks for:

a **statistical model** between the **covariate** $X$ and a **variable** $y$
Probabilistic Downscaling...

... seeks for a **statistical model** between the **covariate** $X$ and a **probabilistic measure** of $y$ - with special emphasis on

$$Q(\tau | X) = y_\tau$$
Probabilistic Downscaling

How to estimate a probabilistic measure of \( y \) given \( X \)?

- Distribution function: \( F(y|X) = \rho \)
- Quantile function: \( Q(\tau|X) = y_\tau \)

Semi-parametric:

Fix threshold \( y \),
estimate 1-\( \rho \) probability of threshold exceedance using logistic regression

\[
\text{Fix probability } \tau=\rho, \text{ estimate } y \text{ quantile using quantile regression}
\]

Parametric:

Distributional assumption:
estimate parameter vector given \( X \)

- Poisson point – GPD process
  estimate parameter \( \mu, \sigma, \xi \) given \( X \)
- POT approach with \( u = u(X) \)
  with parameters \( \sigma_u, \xi \) given \( X \)
Censored Quantile Regression

The **censored** linear model

\[ Y | X = \max(0, \beta^T X + \gamma^T X u) \quad u \sim \text{i.i.d.} \]

the conditional quantile function

\[ Q_y(\tau | X) = \max(0, \beta_{\tau}^T X) \]

**Example**

no censoring:

\[ Q_{\hat{y}}(\tau | X) = \beta_{\tau}^T X \]
Censoring

Conditional quantiles are equivariant to monotone transformations $h(.)$

\[ Q_{h(Y)}(\tau) = h(Q_Y(\tau)) \]

In our case

\[ y = h(\tilde{y}) = \max(0, y) \]

\[ Q_{\tilde{Y}}(\tau \mid X) = \hat{\beta}_T^T X \]

\[ Q_Y(\tau \mid X) = \max(0, \hat{\beta}_T^T X) \]
**Censored Quantile Regression**

The **censored linear model**

\[ Y|X = \max(0, \beta^T X + \gamma^T X u) \quad u \sim \text{i.i.d.} \]

the conditional quantile function

\[ \hat{y}_\tau = Q_y(\tau|X) = \max(0, \beta^T \tau X) \]

**Censored quantile regression** minimizes the **absolute deviation function**

\[ \hat{\beta}_\tau = \arg \min_{\beta_\tau} \sum_n \rho_\tau[y_n - \max(0, \beta^T \tau x_n)] \]

\[ \rho_\tau(u) = \begin{cases} 
\tau u & \text{if } u \geq 0 \\
(\tau - 1)u & \text{if } u < 0 
\end{cases} \]
Peak Over Threshold

Peak over Threshold (POT)

\[
\left\{ Y_i - u \mid Y_i > u \right\}
\]

very large threshold \( u \)

follow a **Generalized Pareto Distribution (GPD)**

Daily precipitation for **Nov-March** (green) and **May-Sept** (red)
Peak Over Threshold

Peak over Threshold (POT)

The distribution of \( Z_i := Y_i - u \mid Y_i > u \) exceedances over large threshold \( u \) are asymptotically distributed following a **Generalized Pareto Distribution (GPD)**

\[
H(z \mid Y_i > u) = \begin{cases} 
1 - (1 + \xi \frac{z}{\sigma_u})^{-1/\xi} & 1 + \xi \frac{z}{\sigma_u} > 0 \quad \xi > 0 \\
1 - \exp\left(-\frac{z}{\sigma_u}\right) & z > 0 \quad \xi = 0 
\end{cases}
\]

two parameters

\( \sigma_u \) scale parameter

\( \xi \) shape parameter
POT with variable threshold

For the model we have 3 parameters

**threshold** $u$ is defined as conditional $\tau$-quantile

$$\hat{u}_\tau = Q_y(\tau_u | X) = \hat{\beta}^T X$$

**scale** and **shape parameter** of GPD

$$\hat{\sigma} = \sigma^T x_n \quad \hat{\xi} = \xi^T x_n$$

Estimation of threshold uses censored quantile regression, conditional GPD parameters of the exceedances are estimated using maximum likelihood method.
POT with variable threshold

The conditional distribution of precipitation is

\[ F(y|X) = H(y - u_x|y > u_x, X) \text{Prob}(y > u_x|X) + \text{Prob}(y \leq u_x|X) \]

\[ = H(u - u_x|y > u_x, X)(1 - \tau_u) + \tau_u = \tau > \tau_u \]

The respective \( \tau \)-is estimated as

\[ \hat{y}_\tau = H^{-1}(\tau|X) = \begin{cases} 
  u_x + \frac{\hat{\sigma}_u}{\hat{\xi}} \left[ (1 - \tilde{\tau})^{-\hat{\xi}} \right] & \hat{\xi} > 0 \\
  u_x + \hat{\sigma}_u \log(1 - \tilde{\tau}) & \hat{\xi} = 0 
\end{cases} \]
Poisson point – GPD process with intensity

\[ \Lambda(A) = (t_2 - t_1) \left[ 1 + \xi \left( \frac{y - u}{\sigma_u} \right)^{-1/\xi} \right] \quad \text{on} \quad A = (t_1, t_2) \times (y, \infty) \]
Poisson Point - GPD

Poisson point – GPD process with intensity

\[ \Lambda(A) = (t_2 - t_1) \left[ 1 + \xi \left( \frac{y-u}{\sigma_u} \right)^{-1/\xi} \right] \text{ on } A = (t_1, t_2) \times (y, \infty) \]

Non-stationarity:
assume parameters to depend linearly on covariate \( X \)

\[ \hat{\mu}_n = \mu^T x_n \quad \hat{\sigma}_n = \sigma^T x_n \quad (\hat{\xi}_n = \xi^T x_n) \]

Estimate parameters using maximum likelihood

Get the conditional extremal quantiles

\[ \hat{y}_\tau = u - \frac{(\hat{\sigma} + \hat{\xi}(u - \hat{\mu}))}{\hat{\xi}} \left( 1 - \left( \frac{-\log \tau}{\hat{\lambda}} \right)^{-\hat{\xi}} \right) \]
Quantile verification score

Quantile regression minimizes the absolute deviation function

\[ QVS = \sum_n \rho_\tau \left[ y_n - \hat{y}_{n,\tau} \right] \]

\[ \rho_\tau(u) = \begin{cases} 
\tau u & \text{if } u \geq 0 \\
(\tau - 1) u & \text{if } u < 0 
\end{cases} \]

Proper scoring rule (Gneiting and Raftery 2005)

censored quantile verification skill score

\[ CQVSS(\tau) = 1 - \frac{CQVS(\tau, \hat{y})}{CQVS(\tau, y_{ref})} \]
Forecast verification

A Scoring rule \( S(P, y)_{y \sim Q} \)

measures the quality or utility of a probabilistic forecast

here in terms of a cost function

Let

\[
E \left[ S(P, y) \right]_{Q} := S(P, Q)
\]

Let \( Q \) be the forecasters best judgement

\( S \) is a proper scoring rule if for all \( P \) and \( Q \)

\[
S(P, Q) \geq S(Q, Q)
\]

strictly proper if

\[
S(P, Q) = S(Q, Q) \iff P = Q
\]

Forecast verification

Compare QR, POT and PP approach
extend censoring to threshold $u$

\[
S_\tau (P_{QR}, y) = \sum_n \rho_\tau \left[ y_n - \max \left( 0, \beta_T^T x_n \right) \right]
\]

\[
S_\tau (P_{POT/PP}, y) = \sum_n \rho_\tau \left[ y_n - \hat{y}_{n,\tau} \right]
\]

\[
S_\tau (P_{ref}, y) = \sum_n \rho_\tau \left[ y_n - Q_y(\tau) \right]
\]

define a quantile verification skill score

\[
QVSS_\tau = 1 - \frac{S_\tau (P_{QR/POT/PP}, y)}{S_\tau (P_{ref}, y)}
\]

cross-validation
Cross validation

Derive forecast using **cross-validation** through separation in **training** period and **target** season

To obtain set of forecasts \( \{ \hat{y}_n \} \) (target season)
we derive estimates of coefficients \( \hat{\beta}_\tau, \hat{\mu}, \hat{\sigma}, \hat{\xi}, \ u_x, \hat{\sigma}_u, \hat{\xi} \)
from set of observations \( \{ \ldots, y_{n'\neq n}, \ldots \} \) (training period)
and covariates \( \{ \ldots, x_{n'\neq n}, \ldots \} \)

The forecast is derived for each target season in the time sequence

\[
\hat{y}_{n, \tau} = \max \left( 0, \beta_T^T x_n \right) \quad \hat{y}_{n, \tau} = Q_{y, \text{POT/PP}} (\tau | x_n)
\]
EVD: Shape Parameter

Winter, $u = 5\text{mm}$
DWD Stations 1948–2004, Winter, PP($u=5.0\text{ mm}$)

Summer, $u = 10\text{mm}$
DWD Stations 1948–2004, Summer, PP($u=10.0\text{ mm}$)
EVD: 20y return values

Winter, $u = 5\text{mm}$

Summer, $u = 10\text{mm}$
Goodness of Fit

u = 5mm

Winter

| Nr. 2 | StN: 80313 |
| Nr. 3 | StN: 80310 |
| Nr. 12 | StN: 54336 |
| Nr. 19 | StN: 71519 |
| Nr. 20 | StN: 71525 |
| Nr. 21 | StN: 70111 |
| Nr. 29 | StN: 73448 |
| Nr. 30 | StN: 71645 |
| Nr. 37 | StN: 90593 |

Summer

| Nr. 2 | StN: 80313 |
| Nr. 3 | StN: 80310 |
| Nr. 12 | StN: 54336 |
| Nr. 19 | StN: 71519 |
| Nr. 20 | StN: 71525 |
| Nr. 21 | StN: 70111 |
| Nr. 29 | StN: 73448 |
| Nr. 30 | StN: 71645 |
| Nr. 37 | StN: 90593 |
QVSS - Winter

\[ \text{POT } \xi = \xi_0 \]
\[ \text{POT } \xi = \xi_0 + \xi_1 x \]
\[ \tau_u = 0.9 \]
QVSS - Winter all Stations

0.99 quantile
DWD Stations 1948–2004, QVSS PP tau=0.990, Winter

0.999 quantile
DWD Stations 1948–2004, QVSS PP tau=0.999, Winter
QVSS - Kempten

![Diagram showing graphs for QVSS with different levels of accuracy (0.9, 0.95, 0.99, 0.995, 0.999)].

- **CQVSS**
  - PP
  - POT
  - QR

Accuracy levels:
- 0.9
- 0.95
- 0.99
- 0.995
- 0.999

Legend for bars:
- Light yellow for 0.9
- Light orange for 0.95
- Orange for 0.99
- Red for 0.995
- Dark red for 0.999
QVSS - Fichtelberg

![Diagram showing QVSS for PP, POT, and QR with different thresholds.]
Fichtelberg August 2002

FICHTELBERG (WEWA)  Summer

precipitation
150
100
50
0

CLIM  1  5  10  15  20  25  30
day of month

99.9%
99.5%
99%
95%
90%
Fichtelberg August 2002

![Graph showing data points and error bars for Fichtelberg August 2002.](image)
Conclusions

Skill
more skill for winter than summer precipitation
(between 20% to 40% in winter, and 10% to 30% in summer)
largest skill obtained for high quantiles

Threshold
largest skill obtained with threshold of about 2 mm

EV parameter
shape parameter is positive in winter and summer
allowing regression for shape parameter induces large error in parameter estimates
-> statistical downscaling model becomes instable
regression for shape parameter increases skill in winter

Outlook
separate convective and stratiform precipitation
estimate quantiles taking into account spacial dependence
References


Koenker, R., 2005: Quantile regression. Cambridge University Press. 349p
