Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression

P. FRIEDERICHS AND A. HENSE
Meteorological Institute, University of Bonn, Bonn, Germany

(Manuscript received 5 July 2006, in final form 15 September 2006)

ABSTRACT

A statistical downscaling approach for extremes using censored quantile regression is presented. Conditional quantiles of station data (e.g., daily precipitation sums) in Germany are estimated by means of the large-scale circulation as represented by the NCEP reanalysis data. It is shown that a mixed discrete–continuous response variable, such as a daily precipitation sum, can be statistically modeled by a censored variable. Furthermore, a conditional quantile skill score is formulated to assess the relative gain of a quantile forecast compared with a reference forecast. Just like multiple regression for expectation values, quantile regression provides a tool to formulate a model output statistics system for extremal quantiles.

1. Introduction

Present day global weather and climate models cannot provide realistic descriptions of local weather and climate variability because of the horizontally restricted resolution (25–100 km at best). Of special interest are extremes of near-surface variables like temperature, wind velocity, or precipitation because of their high impact potential on human activities. Statistical downscaling of extremes intends to provide reliable forecasts of the occurrence probability of extreme events given a model forecast. Forecast models have been validated for their expected behavior, and model output statistics are applied for recalibration and downscaling concerning expectation values. Recently this became a topic of active research regarding extreme events. There is a paradigm shift toward probabilistic weather forecasting. Advances in probabilistic forecasting using ensemble simulations should come along with the development of new statistical modeling techniques that combine dynamical and statistical approaches. This study presents a novel approach for statistical downscaling or recalibration that derives extremal quantile forecasts of precipitation.

In the case of normally distributed response variables (e.g., temperature), only the expectation value and a measure of variance are needed to provide the complete probabilistic behavior of a response variable. The expectation value is estimated by standard linear regression using least squares methods. The variance is derived from the estimated error variance. The underlying assumption is that the error is normally distributed (i.e., that the response variable is conditionally normally distributed). However, variables such as precipitation or wind velocity are clearly not Gaussian. There is no general agreement on which distribution best fits, for example, daily precipitation sums. Another problem with precipitation is its mixed discrete–continuous character.

One method to derive probabilistic forecasts for such a response variable is to estimate quantiles conditional on the large-scale forecast. This technique is referred to as quantile regression (QR). QR was first introduced by Koenker and Bassett (1978). In contrast to standard regression, QR uses least absolute deviations (LAD; $L_1$) to estimate the regression coefficients. The advantage of quantile regression over standard regression is that it only assumes independently identically distributed errors (IID)—an assumption that could even be further relaxed (Koenker 2005).

The first study that applied quantile regression to forecast daily precipitation quantiles was conducted by Bremnes (2004a; in the context of weather forecasting). He used a two-step approach, where first the probability of precipitation and second selected quantiles given...
the occurrence of precipitation were estimated, both conditioned on a forecast of a numerical weather prediction model.

We will present an alternative approach to Bremnes (2004a) where precipitation is represented by a so-called censored variable. Censoring is a term that originates from survival analysis. It arises when the variable of interest is the time until a terminal event (survival time) and the duration of the study is limited in time. If the terminal event occurred some time after the end of the study, it is said to be censored. The censoring line is the end time of the study and represents an upper bound of observation, although the underlying process has no upper bound. In the case of precipitation, the zero precipitation line is assumed to represent a censoring line.

We will show that the approach of Bremnes (2004a) and the censored quantile regression approach both minimize a censored LAD function (Powell 1986). The censored LAD function provides a proper verification score that enables the comparison of forecast approaches for quantiles. A quantile verification (QV) skill score for quantile forecasts can be formulated similarly to the Brier skill score (Brier 1950) for probability forecasts. The QV skill score is not only applicable to a quantile forecast as derived from a QR approach but to any quantile forecast (e.g., as derived from ensemble simulations). Bremnes (2004a) stated the lack of measures of goodness, and no skill score for quantile forecasts is mentioned, for example, in Jolliffe and Stephenson (2003). Although Gneiting and Raftery (2005) present a scoring rule for quantile forecasts, no such score has yet been applied in a meteorological context, at least to our knowledge.

This study is organized as follows. The censored quantile regression and the censored quantile verification (CQV) score are introduced in section 2. Section 3 describes the data used in this study. In section 4, the censored QR applied to forecast conditional quantiles for precipitation at 16 German weather stations is summarized. Section 5 concludes this study.

2. Conditional regression quantiles

Quantile regression is a method used to estimate quantiles of a response variable distribution conditional on a given set of predictors using a linear model. It provides a more complete statistical view than the classical expectation value regression. The quantile regression method was introduced by Koenker and Bassett (1978). Its most common application is in financial mathematics, although recently it has begun to be used in ecology (Koenker and Schorfheide 1994) and meteorology (Bremnes 2004a,b).

We assume a set of random variables \( \{X, Y\} \), where \( Y \in \mathbb{R} \) is the univariate response variable or predictand and \( X \in [1, R^n]^T \) is the conditioning multivariate variable or predictor of the dimension \( q + 1 \). Note that \( X \) has a unity component to account for the intercept \( \beta_{1}^T \) of the regression quantile. The quantile regression for the \( r \) quantile assumes the linear model (Koenker 2005, p. 17):

\[
Y = \beta_r^T X + \gamma^T X \epsilon \in \text{IID.}
\]

Here, \( \beta_r = [\beta_r^{(0)}, \ldots, \beta_r^{(q)}]^T \) are the coefficients of the QR model, \( \epsilon \in \text{IID} \) is an IID distributed error term, and \( \gamma^T X \) is inserted to account for heteroscedasticity (linear-scale effect), that is, root squared variance is a linear function of a transform \( \gamma^T X \), and the super-script \( T \) denotes the transposed vector. Let \( (x_1, \ldots, x_n) \) be the sample of covariates and \( (y_1, \ldots, y_n) \) be the predictand sample. The estimates of the quantile function for a given quantile are defined as

\[
\hat{Q}_r(y|x) = \hat{\beta}_r^T X,
\]

where the vector of the \( r \)th quantile coefficients \( \hat{\beta}_r \) minimizes the piecewise linear LAD function

\[
\hat{\beta}_r = \arg \min_{\beta_r} \sum_{i=1}^n \rho_r(y_i - \beta_r^T x_i),
\]

and \( \rho_r \) denotes the so-called check function with

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

The quantile function has some interesting equivariance properties. Besides scale and shift invariance and equivariance to reparameterization, the regression quantile has the property of equivariance to monotone transformations, notably that for any nondecreasing function \( h(\cdot) \), the equivariance

\[
Q_{hY}(\tau) = h[Q_{rY}(\tau)]
\]

holds (Koenker 2005, p. 41). In other words, the quantile of the transformed variable can be obtained from the transformed quantile of the original variable. In contrast, this is not generally valid for the expectation value. For further insight the reader is referred to Koenker (2005 and references therein). A thorough introduction to quantile regression can also be found in Schulze (2004).

a. Censored quantile regression

The equivariance of the quantile function to monotone transformations (5) is an important property and allows for the formulation of the censored LAD func-
tion (Powell 1986). Let us assume an observable $Y$ with a lower bound of zero (like daily precipitation sums). We further assume that the variate $Y^*$ is linearly generated from a model such as in (1). The censored observable $Y$ is observed through a nondecreasing function

$$Y = h(Y^*) = \max(0, Y^*).$$

Using the equivariance of the quantile function to monotone transformations (5) we obtain

$$Q_{Y^*}(\tau|X) = Q_{\max(0, Y^*)}(\tau|X) = \max[0, Q_{Y^*}(\tau|X)].$$

Hence, for a given quantile $\tau$, the censored regression quantile estimation problem is to minimize the piecewise linear censored LAD function

$$\hat{\beta}_* = \arg\min_{\beta_*} \sum_{i=1}^n \rho_\tau(y_i - \max(0, \beta_*^\top x_i)).$$

### b. Estimation of censored conditional quantiles

The quantile regression is calculated using the R open source statistical computer program (R Development Core Team 2003) and the R quantreg-package of Roger Koenker. He recently included a beta version to calculate regression quantiles for censored variables using the algorithm in Fitzenberger (1997), who adapted the linear programming algorithm for standard QR to censored QR.

Chernozhukov and Hong (2002) proposed an alternative estimation procedure. Here, the estimation procedure is divided into three steps. In step 1, the conditional probability of the occurrence of precipitation $\pi = \text{prob}(Y > 0|X)$ is calculated indicating probability of not censoring. The conditional probability $\pi$ is estimated using a generalized linear model (GLM) with a logit function (Fahrmeir and Tutz 1994). The estimated probability of precipitation is denoted as $\hat{\pi}$. Based on this estimate, a subsample $J_0$ is chosen with $J_0 = \{i: \hat{\pi}_i = \text{prob}(Y > 0|x_i) > 1 - \tau\}$. Step 2 is to calculate an initial estimate of $\beta_*$ using the subsample $J_0$ and standard QR provided by the R quantreg-package. Then an updated subsample with $J_1 = \{i: \hat{\beta}^\top_1 x_i > 0\}$ is selected. Step 3 estimates the three-step estimator $\hat{\beta}_*$ of the censored QR based on the updated subsample $J_1$ using standard QR. An optional step 4 repeats the update of the subsample based on the three-step estimator $\hat{\beta}_*$ with $J_2 = \{i: \hat{\beta}^\top_2 x_i > 0\}$ and updates the estimate $\hat{\beta}_*$.

Both algorithms, the three-step procedure of Chernozhukov and Hong (2002) and the censored QR algorithm provided by the R quantreg-package, give almost identical conditional quantile estimates. However, the three-step procedure turned out to be faster than the censored QR algorithm and is therefore used to derive the results in section 4. Another advantage of the three-step procedure is that it is easily implemented into programs using other programming languages than R where only a standard QR routine exists.

We also applied the procedure described in Brenner (2004a). He estimated the conditional quantile given the occurrence of precipitation. Instead of calculating the conditional $\tau$-quantile based on the whole censored dataset, the estimated conditional $\tau^*$ quantile, with $\tau^* = 1 - (1 - \tau)/\hat{\pi}$, is estimated on the subsample $J_{y>0} = \{i: y_i > 0\}$. This procedure gives estimates $\hat{\beta}_*$, that are very close to those from the other two procedures. However, as $\tau^*$ depends on the covariate (i.e., the conditional probability of the occurrence of precipitation $\pi$), this procedure becomes very computer time consuming, as a $\tau^*$ conditional quantile has to be estimated for each different numerical estimate of the conditional probability $\hat{\pi}$.

### c. Forecast verification

Forecast verification remains an important topic of research in meteorology. Jolliffe and Stephenson (2003) edited a book on different forecast verification approaches. The ultimate goal of forecast verification is to define adequate summary measures (i.e., scoring rules) for the evaluation of a forecast compared with a reference forecast. Scoring rules can be understood in terms of cost functions that a forecast aims at minimizing, or by referring to a Bayesian context as a utility measure that a forecaster wants to maximize. In the following, we define a scoring rule in terms of a cost function. A condition that a scoring rule has to fulfill is that it is proper (Gneiting and Raftery 2005). Assume $S(P, y)$ is a scoring rule, $Q$ is the best distributional forecast, $P$ is an arbitrary distributional forecast, and $y$ is the observed event that materializes. A scoring rule is proper if and only if

$$E[S(P, y)]_{y\sim Q} = E[S(Q, y)]_{y\sim Q} \forall P \neq Q,$$

and strictly proper if, additionally,

$$E[S(P, y)]_{y\sim Q} = E[S(Q, y)]_{y\sim Q} \text{ if and only if } P = Q.$$

$E[S(., y)]_{y\sim Q}$ denotes the expected value of $S(., y)$ when $y$ has a distribution $Q$. Proper scoring rules are proposed and discussed in Gneiting and Raftery (2005).

---

1 The scoring rule can be understood as a distance measure between the forecast and the observations.
The most frequently used score is the Brier score (BS), which is defined for a two-category probability forecast where the categories are separated by a threshold. Given a sample of distributional forecasts \( F_i \) and observations \( y_i, i = 1, \ldots, n \), the BS is estimated as

\[
BS(u) = \frac{1}{n} \sum_{i=1}^{n} [F_i(u) - 1_{y_i \leq u}]^2,
\]

(11)

where \( F_i(u) \) is the conditional forecast distribution function at threshold \( u \) and \( 1_{y_i \leq u} \) is the function that takes a value of 1 if \( y_i \leq u \) and 0 otherwise. An extension of the BS for multiple categories is the ranked probability score (Gneiting and Raftery 2005).

Gneiting and Raftery (2005) proposed a proper scoring rule for quantile forecasts of the form

\[
S(r_i; y^*) = \tau g(r_i) + [g(y^*) - g(r_i)] 1_{y^* > r_i} + f(y^*),
\]

(12)

where \( g \) is a non-decreasing function and \( f \) is arbitrary. With \( g(y^*) = y^* \) and \( f(y^*) = -\tau y^* \), Eq. (12) is identical to the LAD function of the regression quantiles

\[
QVS = \sum_{i=1}^{n} \rho_i (y_i^{*} - \hat{\beta}_{i}' x_i),
\]

(13)

where QVS denotes the quantile verification score for the uncensored variable \( y^* \). Note that Gneiting and Raftery (2005) defined the scoring rule in terms of a reward, thus the sign has to be changed for equivalence.

It is straightforward to define the proper QV score for the precipitation forecast by using the censored LAD function

\[
\text{CQVS}_{\text{cens}} = \sum_{i=1}^{n} \rho_i [y_i - \max(0, \hat{\beta}_{i}' x_i)],
\]

(14)

where CQVS denotes the censored quantile verification (CQV) score for the censored variable \( y \). This equals the proper scoring rule of Gneiting and Raftery (2005) when \( g(y^*) = \max(0, y^*) \) and \( f(y^*) = -\tau \max(0, y^*) \), where \( y = g(y^*) = \max(0, y^*) \) is the censored variable and \( y^* \) is the uncensored variable. By using the (censored) LAD function, we obtain a proper scoring rule that directly applies to estimates of the conditional quantiles of a (censored) forecast variable.

The CQV score is then used to define a skill score that estimates the relative gain of a forecast indexed by “for” over a reference forecast indexed by “ref”. As in most meteorological applications, the reference is taken from the climatological distribution [marginal distribution estimated from the observational sample \( \{y_1, \ldots, y_n\} \)] or a distribution function. Just as the Brier skill score is derived for one specific \( \tau \) quantile can be defined as

\[
\text{CQVS}(\tau) = 1 - \frac{\text{CQVS}_{\text{for} \tau}}{\text{CQVS}_{\text{ref} \tau}}.
\]

(15)

3. Data

A dataset of daily resolution precipitation time series at meteorological stations in Europe that has been compiled for the European Climate Assessment and Dataset (ECAD) project is used as a forecast target (Klein Tank et al. 2002). The procedure is applied to the cold season, November–March (NDJFM), and warm season, May–September (MJJS), daily precipitation sums for 16 stations in Germany for the period from 1948 to 2004. These 16 stations are chosen because the precipitation sums are available for almost all the years during the 57-yr National Centers for Environmental Prediction (NCEP) reanalysis period.

The multivariate covariates are taken from the NCEP reanalysis project (Kalnay et al. 1996). Variables are daily means of relative vorticity \( (\zeta_{500}) \) and vertical velocity \( (w_{850}) \), both at the 850-hPa pressure level, and precipitable water (PWat). The data are available on a 2.5° × 2.5° grid. We have chosen a sector between 5°W and 20°E, and between 42.5° and 60°N, which covers a large part of Europe.

4. Results

a. Forecasting approach

The estimates of the conditional quantiles are derived using cross validation, unless noted otherwise. The daily dataset (i.e., winter or summer daily precipitation sums at one station as the predictand variable and daily NCEP reanalysis over Europe as the multivariate covariate) is separated into a training and target period. The target period comprised the days of one winter or summer season, whereas the training period contains the days of the season from the remaining years. The censored conditional quantile model is estimated using the three-step approach of Chernozhukov and Hong (2002) from the data of the training period, notably the estimates of the coefficients \( \beta \), and the GLM parameter for the probability of censoring \( \hat{\pi} \). The forecasts are derived on the basis of this model fit and the given covariates of the target season. This procedure is repeated for each target season in the 1948–2004 time sequence.

b. Illustration of conditional quantile regression approach

In the following, we want to illustrate the conditional quantile estimation using censored QR. For that pur-
pose, we restrict the covariate to be univariate. The covariate is the NCEP reanalysis gridded PWat field, where the censored QR is applied for each grid point separately. The daily precipitation sums are taken from the Schwerin station for the period from 1948 to 2004.

Figure 1 shows the coefficients and the CQV skill score of a censored QR for the \( \tau = 0.9 \) quantile using PWat at each grid point (i.e., the covariate \( X \) is the daily mean PWat time series at a single grid point). The CQV skill score is calculated using cross validation. Only the \( \hat{\beta}_1 \) (slope) coefficients are shown in Fig. 1; the intercept \( \hat{\beta}_0 \) is not shown. For comparability, the PWat time series at each grid point are normalized.

Similar to standard regression patterns, this field indicates the relative importance and teleconnections in the covariate field with respect to the response variable. In winter, the quantile regression pattern (Fig. 1a) represents a dipole structure with positive coefficients slightly shifted to the north of Schwerin and negative values over southern Europe. It shows that above-normal PWat values near or north of Schwerin as well as drier than normal conditions in the southwest are associated with larger values for the 0.9 quantile of precipitation (i.e., increased probability of large daily precipitation sums). This large-scale teleconnection pattern is also evident in the spatial distribution of the CQV skill score (Fig. 1b). The largest skill of about 15% is obtained with the PWat time series at a grid point northwest of Schwerin at 55°N, 7.5°E. The quantile regression patterns in summer represent the more localized conditions typical for summer rainfall. The coefficients (Fig. 1c) and the CQV skill score (Fig. 1d) are smaller during summer, where the CQV skill score reaches only about 12% at a grid point at 55°N, 12.5°E.

To further illustrate the censored QR approach, we take as covariate the PWat time series at the grid point with the largest CQV skill score, which is at 55°N, 7.5°E in winter and 55°N and 12.5°E in summer. In Fig. 2, the observed daily precipitation sum at Schwerin is plotted against the PWat at the respective grid point (gray asterisks). During the 57 yr, the fraction of rainy days at Schwerin amounts to 0.54 for the winter months and 0.45 for the summer months. Thus, more than 45% (56%) of observed precipitation sums lie on the zero line. The 60%, 75%, 90%, 95%, 97.5%, and 99% conditional quantiles as estimated from the censored QR
Here, \( q_{i.\tau} \) is the number of observations that lie below the conditional \( \tau \) quantile as estimated from the censored QR, then the mean empirical probability of the conditional \( \tau \) quantiles is the proportion of observations below the conditional quantile, \( \eta_i/N \). Here, \( N \) is the total number of observations and \( i = 1, \ldots, N \).

First, the empirical probability was estimated for all forecasts \( q_{i.\tau} \). The empirical probabilities are displayed in Fig. 4 by gray dots. The error bars are derived from 1000 bootstrap samples of the pairs \( (y_i, q_{i.\tau}) \). The reliability diagram shows a significant bias, although small, particularly for the lower quantiles. The bias results from the quantile forecasts that are zero. If the quantile forecast is zero, which means that the estimated probability of censoring is equal to or larger than \( \tau \), then all \( y_i = 0 \) are counted to estimate the empirical probability. This leads to the bias in the empirical probability estimate. We thus excluded the zero forecasts and estimated the empirical probability from \( \eta_i = |\{y_i \leq q_{i.\tau} \} \} | \) and \( N = |\{q_{i.\tau} > 0 \} \} | \), respectively. The bias in the reliability plots (Fig. 4, black circles) now disappears. However, in both cases, the \( \chi^2 \) test for count data (Wilks 1995) does not reject the null hypothesis of a reliable forecast.

Figures 4b,d show the Brier skill score (\( B \)) of the forecast of the occurrence of precipitation using a GLM and the CQV skill score for different quantiles \( \tau \). The Brier skill score amounts to 7% in winter and 3% in summer. It is larger than the CQV skill score for the 0.6 quantile but smaller than the CQV skill score for larger quantiles (\( \tau > 0.7 \)). The CQV skill score is largest for quantiles of \( \tau \approx 0.9 \), which suggests that forecasts for extremal quantiles give larger skill than quantiles near the zero precipitation line, where censoring is more important. The performance of the quantile regression approach declines for small quantiles, where censoring becomes more important (e.g., 0.6 quantile and below). This becomes worse (not shown) for very small quantiles, as most of the time the conditional quantile lies on the censoring line.

To validate the estimates of the conditional \( \tau \) quantiles, we calculate a probability plot (Fig. 4) or reliability diagram, as well as the skill scores. The covariate is again PWat from the grid point with the largest CQV skill score, but now cross validation is applied. The reliability diagram represents the proportion of observations below the conditional quantiles plotted against the forecast probability \( \text{prob}(y \leq q_{i.\tau}) = \tau \). If \( \tau \) is the probability of the conditional quantile and \( \eta_i = |\{y_i \leq q_{i.\tau} \} | \) is the number of observations that lie below the conditional \( \tau \) quantile as estimated from the censored QR, then the mean empirical probability of the conditional \( \tau \) quantiles is the proportion of observations below the conditional quantile, \( \eta_i/N \). Here, \( N \) is the total number of observations and \( i = 1, \ldots, N \).

To validate the estimates of the conditional \( \tau \) quantiles, we calculate a probability plot (Fig. 4) or reliability diagram, as well as the skill scores. The covariate is again PWat from the grid point with the largest CQV skill score, but now cross validation is applied. The reliability diagram represents the proportion of observations below the conditional quantiles plotted against the forecast probability \( \text{prob}(y \leq q_{i.\tau}) = \tau \). If \( \tau \) is the probability of the conditional quantile and \( \eta_i = |\{y_i \leq q_{i.\tau} \} | \) is the number of observations that lie below the conditional \( \tau \) quantile as estimated from the censored QR, then the mean empirical probability of the conditional \( \tau \) quantiles is the proportion of observations below the conditional quantile, \( \eta_i/N \). Here, \( N \) is the total number of observations and \( i = 1, \ldots, N \).

First, the empirical probability was estimated for all forecasts \( q_{i.\tau} \). The empirical probabilities are displayed in Fig. 4 by gray dots. The error bars are derived from 1000 bootstrap samples of the pairs \( (y_i, q_{i.\tau}) \). The reliability diagram shows a significant bias, although small, particularly for the lower quantiles. The bias results from the quantile forecasts that are zero. If the quantile forecast is zero, which means that the estimated probability of censoring is equal to or larger than \( \tau \), then all \( y_i = 0 \) are counted to estimate the empirical probability. This leads to the bias in the empirical probability estimate. We thus excluded the zero forecasts and estimated the empirical probability from \( \eta_i = |\{y_i \leq q_{i.\tau} \} \} | \) and \( N = |\{q_{i.\tau} > 0 \} \} | \), respectively. The bias in the reliability plots (Fig. 4, black circles) now disappears. However, in both cases, the \( \chi^2 \) test for count data (Wilks 1995) does not reject the null hypothesis of a reliable forecast.

Figures 4b,d show the Brier skill score (\( B \)) of the forecast of the occurrence of precipitation using a GLM and the CQV skill score for different quantiles \( \tau \). The Brier skill score amounts to 7% in winter and 3% in summer. It is larger than the CQV skill score for the 0.6 quantile but smaller than the CQV skill score for larger quantiles (\( \tau > 0.7 \)). The CQV skill score is largest for quantiles of \( \tau \approx 0.9 \), which suggests that forecasts for extremal quantiles give larger skill than quantiles near the zero precipitation line, where censoring is more important. The performance of the quantile regression approach declines for small quantiles, where censoring becomes more important (e.g., 0.6 quantile and below). This becomes worse (not shown) for very small quantiles, as most of the time the conditional quantile lies on the censoring line.

To assess whether the linear quantiles provide a good fit of the data, we also considered a quadratic dependency on PWat. Neither in winter nor in summer is a...
quadratic dependency of precipitation on PWat found. The respective $\hat{\beta}_2^{(2)}$ coefficient is not significantly different from zero. Furthermore, a second-degree polynomial fit of the conditional quantile does not significantly increase the CQV skill score. Another way to detect inadequate fits is to consider the reliability conditioned on the value of the quantiles. For that purpose we divided the dataset into parts where the forecast quantiles lie below and above the median of forecast quantile distribution. For both parts, the reliability plot showed no deviations from the empirical distribution (not shown).

c. Results

Up to now, the results aimed at representing and validating the censored quantile regression approach. We now want to examine potential forecast skills for the German weather stations as provided by the ECAD using as covariate a combined vector of $\zeta_{850}$, $\omega_{850}$, and PWat. As described in the data section, the multivariate covariates are taken from the NCEP reanalysis dataset.

First, we have to quantify the optimal number of spatial degrees of freedom representing the covariate, as prior to model fitting, a reduction of spatial degrees of freedom of the covariate is necessary. Here we use common principal component analysis. The multivariate covariate that consists of about $10 \times 18$ grid point time series over the European sector for 3 meteorological variables ($\zeta_{850}$, $\omega_{850}$, and PWat) is projected onto its leading empirical orthogonal functions (EOFs). To give equal weight to the different meteorological variables, the time series are weighted by the inverse of the spatial mean temporal standard deviation of the variable. This is necessary because of the strong discrepancy between the standard deviations, which is on the order of $10^{-4}$ s$^{-1}$ for $\zeta_{850}$, $10^{-1}$ m s$^{-1}$ for $\omega_{850}$, and 50 kg m$^{-2}$ for PWat.

Figure 5 shows the CQV skill score for varying numbers of EOFs ranging from 3 to 23. The skill scores are calculated from the cross-validated forecasts. In winter, the CQV skill score increases with increasing numbers of EOFs up to about 18 EOFs. Afterward, the CQV skill score remains constant. In summer, the maximum CQV skill score is reached with about 19–20 EOFs. Similar results are obtained for the other stations. Accordingly, we think that it is convenient to represent the multivariate covariate by 20 EOFs. In winter, the first 20 EOFs explain about 90% of the total variance of the combined NCEP field; in summer this amounts to 86%.

As the conditional quantiles are estimated independently for the different levels of $\tau$, it may occur that the quantile estimates cross (i.e., the conditional 0.95
quantile is larger than the 0.99 quantile) and constitute invalid distributions. We investigated the number of crossing quantiles for each station. When using 20 EOFs, a crossing of the conditional quantiles occurs during much less than 0.1% of the approximately 8600 days. For many stations, the number of crossing quantile estimates is zero. However, when too many EOFs were included, some stations showed a considerable number of crossing events. Crossing almost disappears when the number of EOFs was reduced. Besides the CQV skill score, oversampling could thus be detected when checking the occurrence of crossing quantile estimates. In cases where the crossing of quantiles remains a problem, it might be necessary to include constraints into the estimation process of the $\beta$ coefficients that avoid crossing (Bremnes 2004a).

The following results are for a censored QR with a covariate that consists of the first 20 EOFs of the combined vector of $\xi_{850}$, $\omega_{850}$, and PWat. Figure 6 shows the probability plot and the skill scores for Schwerin. The bias, as seen in the reliability diagrams (Figs. 6 a,c) when the empirical probability is estimated including the zero quantile forecasts (Figs. 6 a,c, gray dots), is more pronounced than in Fig. 4, particularly for the lower quantiles and the summer forecast. The $\chi^2$ test for count data rejects the null hypothesis of reliability. However, if the reliability is assessed excluding the zero quantile forecasts, the bias disappears (Figs. 6a,c, circles). The CQV skill score reaches between 30% for the 0.6 quantile and 41% for the 0.9 quantile in winter. In summer, the CQV skill score amounts to 24% for the 0.9 quantile. The highest skill is obtained for the 0.9 quantile in summer as well as in winter. The Brier skill score, for the probability forecast using the GLM for precipitation yes or no, amounts to 36% in winter as well as in summer. Thus, the largest skill in summer is obtained for the probability forecast of the occurrence of precipitation.

The CQV skill score for 16 German stations and the quantiles $\{0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99\}$ are displayed in Fig. 7. The sampling error of the CQV skill score is estimated as the 97.5% (2.5%) quantile of a bootstrap
sample of size 1000. Almost all stations show much larger skill scores for the winter compared with the summer seasons. Furthermore, maximum CQV skill score is generally obtained for the 0.9 quantile. However, Dresden shows particularly large prediction skill for the extremal quantile ($\tau = 0.99$). Note that the sampling error becomes large for the 0.99 quantile. There are significant differences in the CQV skill score between the different stations. Two reasons seem possible for the strong differences in skill: either some physical processes responsible for precipitation at a station are better captured by the NCEP reanalysis than others or the data quality strongly differs from station to station. The latter could be the case for stations that are very close, like Potsdam and Berlin (distance of approximately 40 km), which show large differences in the CQV skill score, especially for the summer seasons.

We want to continue with two examples in actual downscaling. In December 1993, a severe flooding occurred in northwestern Europe. Cologne and Bonn experienced the first of the two worst floods in the twentieth century (the second was in January 1995). The flood was primarily caused by excessive precipitation over large parts of western and central Europe, which was related to an active western circulation (Engel et al. 1994; Caspary 1996). The second example is the severe flood that occurred in central and eastern Europe in August 2002 (Rudolf and Rapp 2003). Heavy long-lasting rainfall induced flooding of major rivers in the Czech Republic, Austria, and Germany. The heavy rainfall was due to a low pressure system that passed across Europe and a depression that occurred slightly earlier south of the Alps.

Figure 8 shows the observed daily precipitation sums and the QR forecasts for Frankfurt in December 1993. The observations are indicated by the black dots. On 19 December, about 16 mm had already fallen prior to the heavy rainfall on 20 December of about 39 mm. Furthermore, over 60 mm of precipitation was observed during one week in the first half of December 1993. The total precipitation sum in December 1993 amounted to 175 mm. The forecast approach uses 20 EOFs to represent the combined field of $z_{850}$, $\omega_{850}$ and PWat of the NCEP reanalysis. First, a forecast for the probability of precipitation is derived using a GLM with a logit function. It is represented by gray bars in Fig. 8. A box and
whisker diagram represents the conditional 0.9, 0.95, and 0.99 quantiles given the NCEP data field of that respective date. Except during the first week of December, the probability of rainfall is large at about 0.8 for most days against a climatological probability of 0.45. On 20 December, the forecast probability of rain is 0.95. The conditional 0.9 quantile is about 17 mm, and the 0.99 quantile is about 23 mm. While the precipitation sum on 19 December lies between the 0.95- and 0.99-quantile forecast, the rainfall on 20 December greatly exceeded the 0.99-quantile forecast, which estimates the probability for precipitation above 23 mm to be 0.01. A climatological forecast would only give 4.5 mm for the 0.9 quantile and 11.8 mm for the 0.99 quantile.

Figure 9 shows the same forecast for Dresden during August 2002. In Dresden, more than 180 mm of precipitation were measured from 12 to 13 August. On 12 August, the rainfall amount exceeded 120 mm. The forecast probability for precipitation for 12 August amounts to 0.99 versus 0.45 for climatology, and there was a 10 % chance that the precipitation exceeded 34 mm day$^{-1}$. The conditional 0.99 quantile amounts to 64 mm on 12 August and remains high with 42 mm the succeeding day. A climatological forecast predicts only 6.3 mm for the 0.9 quantile and 21.1 mm for the 0.99 quantile.

d. Influence of the length of the training dataset

Hitherto, the training of the quantile regression model relies on daily values from a dataset of 56 yr. It is implicitly assumed that the relationship between the covariates and the daily rainfall quantiles remains constant and that the data are homogeneous. Furthermore, in operational forecasting typically only a few years of homogeneous forecast data are available. We want to assess the performance of the quantile regression approach in situations where the available training dataset has only a length of 5 and 10 yr. The target period again comprised the days of one winter or summer season. The training period contains the days of the respective season from the 5 (10) preceding years. If not enough preceding years are available, the training dataset is filled with the days from the following years. The number of EOFs for the 5- and 10-yr training approach was set to 13, when, in most cases, the CQV skill score reaches a maximum. However, the crossing of the conditional quantile estimates becomes much more frequent. Crossing quantiles occur during 25 % of the 8600 days for a 5-yr training period and about 10 % for the
10-yr training period, and the number increases strongly with the number of EOFs.

We choose as examples four stations—Bamberg, Dresden, Schwerin, and Munich. The other stations behave very similarly. The CQV skill scores as obtained using the different training periods are represented in Fig. 10. For all stations, the CQV skill scores for the 0.99 quantile are significantly smaller when using only 5 yr of training data. For the other quantiles there is no consistent performance against the length of the training period. The CQV skill scores remain almost constant for Bamberg, while it significantly decreases for Dresden in summer and for Schwerin in summer and winter. For Munich, the CQV skill scores using the short training periods are significantly larger than using all data for the training. This could indicate inhomogeneities or nonstationarity in the relationship between the covariates and the precipitation quantile.

5. Conclusions

We presented a novel statistical downscaling approach to estimate conditional quantiles of precipitation at a single station given the large-scale circulation as represented by the NCEP reanalysis data. The approach uses QR, which was first introduced by Koenker and Bassett (1978). To our knowledge, the first and hitherto only application in a meteorological context was carried out by Bremnes (2004a).

Two extensions of the QR approach mark this work. 1) This study shows that a mixed discrete–continuous response variable such as daily precipitation sums can be statistically modeled by a censored variable. The conditional quantiles are then calculated using the censored QR approach, which provides a much faster procedure to estimate the model parameters than the two-step approach used in Bremnes (2004a), especially when using cross validation. 2) The censored QR that aims at minimizing a censored LAD function provides a natural basis for a proper scoring rule. This scoring rule allows the CQV skill score to be formulated, which assesses the relative gain of a quantile forecast compared with a reference forecast. This skill score for quantile forecasts is the analog to the Brier skill score for probability forecasts. To our knowledge, no such
type of score has ever been used in meteorological or climate applications.

Although the assumption of linear dependency seems reasonable in our example, and the crossing of the conditional quantiles constitutes no problem, there might be cases where the linear assumption is inconvenient or crossing quantiles indeed constitute a problem. In those cases, local QR as applied by Bremnes (2004a) combined with the three-step censoring approach should be considered.

We illustrate the censored QR approach by means of precipitation forecasts at 16 German weather stations. The multivariate covariate is composed of the EOFs of a combined vector of relative vorticity and vertical velocity at 850 hPa and precipitable water over Europe. Of special interest are the extremal quantiles,
such as the 0.9, 0.95, and 0.99 quantiles of precipitation. The largest CQV skill score was generally obtained for the 0.9 quantile. Although the sampling error for the 0.99 quantile becomes significantly larger, the 0.99 quantile also shows comparable skill. The CQV skill score of the 0.9-quantile forecast ranges between 23% and 41% in winter and between 17% and 28% in summer. Large differences in the CQV skill score are found between the German stations, even between those that are very close to each other. Only a better observational network would permit us to distinguish regional differences from differences due to the lack of data quality.

It is shown that the conditional quantile forecasts are sensitive enough to assess the probability of extreme events. In both examples (Frankfurt in December 1993 and Dresden in August 2002), the 0.99-quantile forecast is anomalously high. However, the measured rainfall on 20 December in Frankfurt as well as on 12 August in Dresden greatly exceeded the 0.99-quantile forecast.

Better skill could be obtained by choosing more suitable forecast variables. Particularly during summer, where convective precipitation prevails, a variable representing the convectively available potential energy would certainly increase the forecast skill (A. Mathes 2006, personal communication). Furthermore, the resolution of the NCEP reanalysis data with a $2.5^\circ$ resolution is relatively coarse. A higher resolution of the covariate would probably provide more skill.

The quantile regression approach is expected to give reliable forecasts even when the training dataset comprises only a sequence of 5 yr of daily data. In our case, 5 yr of training data seems too short to obtain good forecast skill for very high precipitation quantiles such as the 0.9, 0.95, and 0.99 quantiles of precipitation. The largest CQV skill score was generally obtained for the 0.9 quantile. Although the sampling error for the 0.99 quantile becomes significantly larger, the 0.99 quantile also shows comparable skill. The CQV skill score of the 0.9-quantile forecast ranges between 23% and 41% in winter and between 17% and 28% in summer. Large differences in the CQV skill score are found between the German stations, even between those that are very close to each other. Only a better observational network would permit us to distinguish regional differences from differences due to the lack of data quality.

It is shown that the conditional quantile forecasts are sensitive enough to assess the probability of extreme events. In both examples (Frankfurt in December 1993 and Dresden in August 2002), the 0.99-quantile forecast is anomalously high. However, the measured rainfall on 20 December in Frankfurt as well as on 12 August in Dresden greatly exceeded the 0.99-quantile forecast.

Better skill could be obtained by choosing more suitable forecast variables. Particularly during summer, where convective precipitation prevails, a variable representing the convectively available potential energy would certainly increase the forecast skill (A. Mathes 2006, personal communication). Furthermore, the resolution of the NCEP reanalysis data with a $2.5^\circ$ resolution is relatively coarse. A higher resolution of the covariate would probably provide more skill.

The quantile regression approach is expected to give reliable forecasts even when the training dataset comprises only a sequence of 5 yr of daily data. In our case, 5 yr of training data seems too short to obtain good forecast skill for very high precipitation quantiles such
as the 0.99 quantile, and the crossing of the conditional quantiles becomes an issue when the training period is short.

Technically, the QR is relatively easy to apply, as programs are provided, amongst others, by Roger Koenker. The three-step censored QR after Chernozhukov and Hong (2002) uses such a standard QR procedure to compute the censored QR. The CQV skill score allows one to compare the relative gain of different forecast approaches. Furthermore, the CQV skill score can be applied to evaluate any quantile forecast, whether derived from a single deterministic model simulation or from an ensemble of simulations. It enables the forecaster to search for suitable predictor variables. Quantile regression can be used to derive recalibrated forecasts for extremal quantiles, exactly as model output statistics use linear regression to estimate a linear recalibration of the expectational model forecasts. dall 2001) must be used.

Acknowledgments. Special thanks are due to Tilman Gneiting, Ian Jolliffe, and the anonymous reviewers, who gave very helpful and detailed comments.

REFERENCES


