Adv. Stat. Clim. Meteorol. Oceanogr., 2, 39–47, 2016
www.adv-stat-clim-meteorol-oceanogr.net/2/39/2016/
doi:10.5194/ascmo-2-39-2016

ASCMO
Open Access

Calibrating regionally downscaled precipitation over Norway through quantile-based approaches

David Bolin[1], Arnoldo Frigessi[2,4], Peter Guttorp[3,4], Ola Haug[4], Elisabeth Orskaug[4], Ida Scheel[5], and Jonas Wallin[1]

[1]Dept. of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden
[2]Dept. of Biostatistics, University of Oslo, Oslo, Norway
[3]Dept. of Statistics, University of Washington, Seattle, WA, USA
[4]Norwegian Computing Center, Oslo, Norway
[5]Dept. of Mathematics, University of Oslo, Oslo, Norway

Correspondence to: Ola Haug (ola.haug@nr.no)

Received: 22 December 2015 – Revised: 4 April 2016 – Accepted: 12 May 2016 – Published: 9 June 2016

Abstract. Dynamical downscaling of earth system models is intended to produce high-resolution climate information at regional to local scales. Current models, while adequate for describing temperature distributions at relatively small scales, struggle when it comes to describing precipitation distributions. In order to better match the distribution of observed precipitation over Norway, we consider approaches to statistical adjustment of the output from a regional climate model when forced with ERA-40 reanalysis boundary conditions. As a second step, we try to correct downscalings of historical climate model runs using these transformations built from downscaled ERA-40 data. Unless such calibrations are successful, it is difficult to argue that scenario-based downscaled climate projections are realistic and useful for decision makers. We study both full quantile calibrations and several different methods that correct individual quantiles separately using random field models. Results based on cross-validation show that while a full quantile calibration is not very effective in this case, one can correct individual quantiles satisfactorily if the spatial structure in the data are accounted for. Interestingly, different methods are favoured depending on whether ERA-40 data or historical climate model runs are adjusted.

1 Introduction

The intensification of climate research over the past decade produces a steadily increasing number of data sets combining different global circulation or earth system models, CO_2 emissions scenarios and downscaling techniques. Turning future projections into robust and reliable information available at a local scale is imperative for the successful modelling of impacts of climate change in nature and society. The comprehensive financial and safeguarding challenges of mitigation and adaptation call for thorough validation, improvement and extensions of current downscaling techniques.

The comparison of climate models to weather data raises interesting statistical problems. For a statistician, the most natural definition of the climate is that it is the distribution of weather (and other earth system variables) over multi-decadal timescales (Smith et al., 2010; Guttorp, 2014). A climate model (general circulation model or more generally earth system model) describes the distribution of observable variables based on physical principles. Because some of the processes (e.g. convection, clouds) occur on scales smaller than the large grid squares needed to approximate a solution to the Navier–Stokes equations, such processes are often calculated using simple approximations (or parameterizations).

A multitude of models have emerged for projection of future climate change at different spatial (and temporal) scales. Essential in the process of going from the coarse resolution of the global models to finer spatial scales are the regional climate models (RCMs). Such models propagate information from a coarse-scale model along the boundary of a higher-resolution area of interest, using a more detailed terrain description, model solutions using finer resolution, and

improved physical process parameterizations. The boundary conditions may be computed either from a global weather model forced with updated historical observations to calculate consistently the state of the atmosphere (reanalysis), or from a global climate model. A regional model using reanalysis boundary conditions is sometimes said to be run in "weather forecasting mode", and is the closest one can hope to get to observed weather using a regional climate model. Maraun et al. (2010) described approaches to downscaling precipitation.

One major purpose of regional climate models is to give end users such as stakeholders and decision makers a representation, preferably a reliable projection, at a practically useful spatio-temporal scale, of future weather. In the insurance industry, for instance, the interest lies in high precipitation projections under various possible future scenarios to assess the changing risk of damages to buildings or flooding (Scheel and Hinnerichsen, 2012). Typically, scenario runs are built from the regional model forced by global coupled ocean–atmosphere or earth system model runs. The question then becomes how reliable these regional models are, at the scale needed by the actual effect study. For example, to understand patterns of risk for the insurance of buildings, precipitation at meso- to local-scale level is needed.

Orskaug et al. (2011) compared precipitation from the HIRHAM regional model (Bjørge and Haugen, 1998), run over Europe and forced with ERA-40 reanalysis (Uppala et al., 2005) boundary conditions, to a gridded precipitation product for Norway (Jansson et al., 2007) on a $25 \times 25\,\mathrm{km}^2$ scale. They used a variety of statistical measures for comparing the two data sets. The regional model output was found to describe low levels of precipitation fairly well, but failed to reproduce large quantities. Maule et al. (2013) found that most of the regional models in ENSEMBLES give fairly accurate descriptions of drought indices and other functions of low precipitation regimes. These findings, together with the need for representative scenarios called for by most impact studies, serve as a motivation for improving the local-scale description of extreme future climate precipitation.

It has long been understood that regional models tend to be regionally biased in terms of precipitation (e.g. Christensen et al., 2008; Monjo et al., 2014; Mishra et al., 2014). Bias correction is an approach that attempts to adjust (statistically or otherwise) the climate model output (regional or global) to make it closer to observed data for historical runs. The idea is then that applying this bias correction to future simulations should also provide more realistic projections. Kerkhoff et al. (2014) develop a framework for assessing the bias correction, and the assumptions needed to apply such corrections to projections. They focus on temperature data, and can therefore assume normal distributions, which is not appropriate for precipitation.

There are a variety of bias correction methods in the literature (Maraun et al., 2010, contains a review). The simplest is a multiplicative correction to make the empirical means of data and RCM output agree (Lenderink et al., 2007). Some authors (e.g. Schmidli et al. (2007) prefer to make the correction only to the mean of precipitation on rainy days. An intermediate approach adjusts the coefficient of variation of data and model output (Teutschben and Seibert, 2012). More advanced methods try to match quantiles, either by fitting gamma distributions (with point mass at zero) to data and models (Piani et al., 2010) or non-parametrically using full quantile mappings (Themeßl et al., 2011). The bias corrections are typically done grid square by grid square, without an explicit spatial model for between grid square dependence. Quantile corrections (or smoothed estimates thereof) typically are found superior in comparative assessments (Gudmundsson et al., 2012; Räty et al., 2014; Fang et al., 2015).

In this paper we consider approaches to statistical adjustment of the regional model output, obtaining a calibrated product that is closer in distribution to the observed data than the original output. We first investigate the Doksum shift function (Doksum, 1974), which makes a full quantile calibration, as the basic tool for adjustment. Next, we restrict ourselves to less ambitious models that correct individual quantiles separately. Considering gridded data products covering Norway, we build transformations either separately for each grid cell, or via models that incorporate some kind of spatial structure. The models are fitted to a training set of downscaled ERA-40 data, and then used to correct downscaled ERA-40 on a test set. We also try to correct downscalings of historical climate model runs using the same transformations built on downscaled ERA-40 data. Unless such calibrations are successful, it is difficult to argue that scenario-based downscaled climate projections are realistic and useful for decision makers.

The paper continues as follows. In Sect. 2 we present the various data sets used in the analysis. Section 3 deals with using the shift function to do full quantile bias correction, while Sect. 4 focuses on bias correction of individual quantiles. Section 5 discusses the potential use of the methodology in assessment and uncertainty quantification of regional climate models.

All code needed to run the analysis on the data are found at Bolin et al. (2016).

2 Data

The data used in this study constitute 40 years of daily precipitation values for the Norwegian mainland, covering the period 1961 to 2000. The data set is twofold: one part consists of dynamically downscaled model data (ERA-40 reanalysis and climate model), and the other is a gridded product based on in situ observations. A more thorough description of the data are given in Orskaug et al. (2011).

2.1 ERA-40 reanalysis

Reanalysis data express the best, physically consistent, estimate available for the historical state of the atmosphere. They are formed in retrospect from feeding various sources of past meteorological observations into a current meteorological forecast model. ERA-40 reanalysis data (Uppala et al., 2005) are a product of the European Centre for Medium-Range Weather Forecasts in the UK.

Downscaled ERA-40 data are collected from the ENSEMBLES project website (http://ensemblesrt3.dmi.dk) (Christensen et al., 2010). Gridded large-scale ERA-40 data along the boundary of an integration area covering most of Europe are dynamically downscaled to weather variables on a grid with a spatial resolution of $25 \times 25\,\mathrm{km}^2$, which amounts to 777 grid cells covering the Norwegian mainland. The downscaled ERA-40 reanalysis data will be referred to as dERA40 in this paper. The downscaling is done by the Norwegian Meteorological Institute using their HIRHAM Regional Climate Model (Bjørge and Haugen, 1998).

2.2 Observation-based data

Precipitation is measured daily at stations irregularly distributed across Norway. Based on all observations of precipitation available at every time step, high-resolution precipitation grids ($1 \times 1\,\mathrm{km}^2$) are estimated applying a Delaunay triangulation (Jansson et al., 2007). The interpolated precipitation values are adjusted locally by taking the deviations between triangulated station elevations and ground heights as given by a detailed terrain model into account. Also, prior to the interpolation, observed precipitation is corrected for exposure-dependent undercatch due to wind loss (Førland et al., 1996).

In order to compare the two data sets, the $1 \times 1\,\mathrm{km}^2$ observation grid was aggregated into the larger $25 \times 25\,\mathrm{km}^2$ grid of dERA40. This was obtained by collecting all $1 \times 1\,\mathrm{km}^2$ grid cells with centre points within a dERA40 cell, and taking their average as a representation of the measured precipitation inside that grid cell. We use the abbreviation OBS for this data set.

2.3 Climate model data

The global Bergen Climate Model, BCM, data set (Furevik et al., 2003) is downscaled by the Norwegian Meteorological Institute using the same HIRHAM Regional Climate Model as for dERA40 and also collected from the ENSEMBLES project website. This downscaled climate data set has the same spatial resolution of $25 \times 25\,\mathrm{km}^2$. The downscaled BCM climate model data set will be referred to as dBCM in this paper.

3 Full quantile calibration of Norwegian precipitation

The results of the evaluation of the regional model (Orskaug et al., 2011) underlines the need for enhanced climate projections at a local scale. Discrepancies between the distributions of observed and downscaled precipitation exist for the whole range of data, suggesting that a full quantile calibration function is needed. In Sect. 3.1 we address this issue using a calibration that will make the model data distribution closer to that of the observed data.

3.1 Distributional calibration using Doksum's shift

We characterize the transfer function between two distribution functions, in our case those of a model and of observations, using Doksum's shift function (Doksum, 1974). To define this function, consider data from two distributions, F and G, and let $\Delta(x) = G^{-1}(F(x)) - x$. If $X \sim F$ (i.e. X is a random variable with cumulative distribution function F), it is easy to see that $X + \Delta(X) \sim G$. In other words, $\Delta(x)$ measures how much the distribution F needs to be shifted at a value x in order to coincide with the distribution G. The shift function Δ can be estimated using empirical distribution functions for F and G:

$$\widehat{\Delta}(x) = \widehat{G}^{-1}(\widehat{F}(x)) - x,$$

where $\widehat{F}$ and $\widehat{G}$ are the empirical cumulative distribution functions of F and G, respectively. The empirical cumulative distribution function is a step function:

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le t),$$

where $I(A)$ is the indicator of event A, and $(x_1, \ldots, x_n)$ are observations of independent and identically distributed real random variables distributed according to F. We follow the standard statistical notation where X_i stands for a random variable and x_i for its observed value.

If the shift function is constant, it means that there is only a difference in location between the two distributions (and particularly if that constant equals zero there is no difference between the distributions). If it is linear, a location-scale transformation is implied.

Assume next that a region is divided into S grid cells. For grid cell $i, i = 1, \ldots, S$, let X_i denote downscaled model precipitation and Y_i observations. Let F_i be the cumulative distribution function of X_i and G_i that of Y_i. Assume further that we have downscaled model output x_{it} and observations y_{it} for days $t = 1, \ldots, T$ (using a common T implies no loss of generality; should the number of data points for $\widehat{F}$ and $\widehat{G}$ rather be $T_{\widehat{F}}$ and $T_{\widehat{G}}$, respectively, those are used instead). Our interest lies in distributional coherence between the downscaled model data and the observations, rather than daily correspondence between x_{it} and y_{it}.

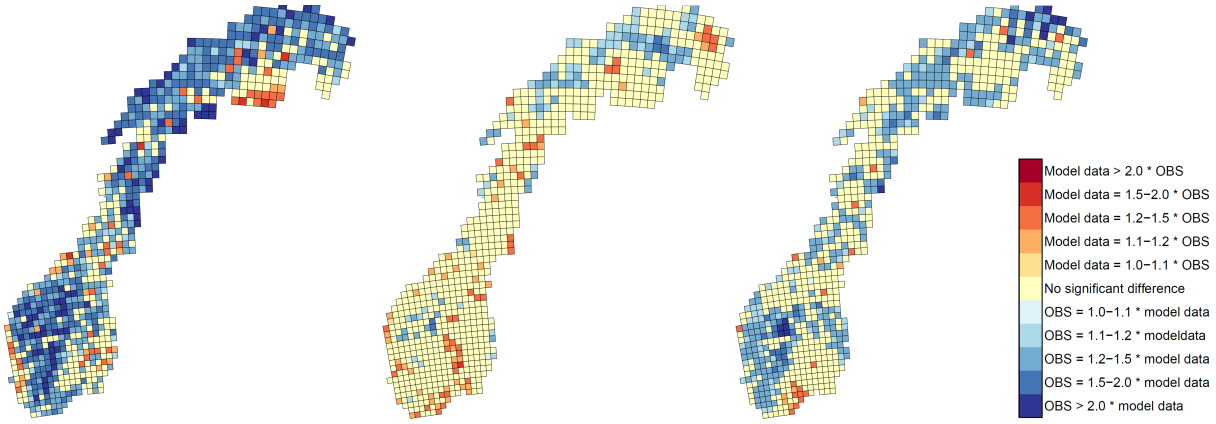

Figure 1. Fisher tests of the 95 % quantile of the winter season uncalibrated dERA40 test data (left panel), the calibrated dERA40 test data (middle panel), and the calibrated dBCM test data (right panel). The plots show significance level $\alpha = 5\,\%$.

Calibration of a new value $x_{it'}^{\text{uncal}}$ drawn from the same distribution $\widehat{F}_i$ but with $t' \notin 1,\ldots,T$ is done by adding its Doksum shift,

$$\begin{aligned} x_{it'}^{\text{cal}} &= x_{it'}^{\text{uncal}} + \widehat{\Delta}_i(x_{it'}^{\text{uncal}}) \\ &= x_{it'}^{\text{uncal}} + \widehat{G}_i^{-1}(\widehat{F}_i(x_{it'}^{\text{uncal}})) - x_{it'}^{\text{uncal}} \\ &= \widehat{G}_i^{-1}(\widehat{F}_i(x_{it'}^{\text{uncal}})), \end{aligned} \tag{1}$$

showing that this calibration is indeed a full quantile calibration.

3.2 Transferability of calibration

Assume that we want to apply the calibration in Eq. (1) to data from another data set, i.e. to $z_{it'}^H$ not necessarily distributed according to F and where t' may or may not overlap with $1,\ldots,T$. A typical example would be extending the calibration established for a re-analysis to a historical climate model run. As before, we use data from F and G to calculate empirical distributions $\widehat{F}_i$ and $\widehat{G}_i$ respectively, where $\widehat{F}_i$ is estimated from x_{it}, $t = 1,\ldots,T$ and $\widehat{G}_i$ from y_{it}, $t = 1,\ldots,T$. We then correct the new data set, $z_{it'}^H$, by

$$\begin{aligned} z_{it'}^{\text{cal}} &= z_{it'}^H + \widehat{\Delta}_i(z_{it'}^H) \\ &= z_{it'}^H + \widehat{G}_i^{-1}(\widehat{F}_i(z_{it'}^H)) - z_{it'}^H \\ &= \widehat{G}_i^{-1}(\widehat{F}_i(z_{it'}^H)). \end{aligned}$$

3.3 Shift function calibration results

In Orskaug et al. (2011) it was shown using several criteria that the dERA40 and OBS data sets lack agreement. A detailed comparison of specific local features showed that the global disagreement was due to poor agreement for high quantiles. As mentioned in Orskaug et al. (2011) the day-by-day correlation is partly lost when downscaling the ERA-40 data; hence we compare distributions rather than using day-by-day test measures. We test the calibration models described in Sects. 3.1 and 3.2 by considering different seasons separately. The seasons used are winter (December to February), spring (March to May), summer (June to August) and autumn (September to November).

In our current setup, the dERA40 and OBS data are further divided into a training set and a test set. The training set is used to fit the calibration model. The transfer function thus obtained is applied to dERA40 data for the test period, which then is compared to observations for the test period. Here the training data are chosen to be the first 80 % of the total data, i.e. the years from 1961 to 1992. The test data are chosen to be the last 20 % of the data, i.e. the years from 1993 to 2000.

The dERA40 data are calibrated using Doksum's shift function as described in Sect. 3.1. In particular, for a specific x_{it}^{uncal} from the test data set, its calibrated value x_{it}^{cal} is calculated from Eq. (1) where $\widehat{F}$ and $\widehat{G}$ both are estimated based on the training data.

Assuming independence between the test statistics for different grid squares, if all null hypotheses are true, we would expect about 39 spurious significances at 95 % confidence level in a plot with 777 grid cells. We have carried out the same kind of comparisons as in Orskaug et al. (2011), but here we only report the Fisher test of the 95th percentile. Figure 1 shows substantial amounts of rejections (74 %) in the uncalibrated dERA40, with an improvement (24 % rejections) for the calibrated data, and a deterioration (48 % rejections) for the downscaled Bergen climate model. Things are worse for the Kolmogorov–Smirnov test, in particular for the climate model data (77, 18 and 79 %, respectively). Since we are estimating the calibration from the training data, we do not expect to get only 5 % rejections in the test set. Furthermore, spatial dependence also affects the rejection rates.

The main reason for the difficulty of making a full quantile calibration is that the bulk of the distribution is concentrated around very small precipitation values, and the Kolmogorov–Smirnov statistic tends to focus on these well-estimated parts

of the distribution, where very small differences in amounts are the reason for rejection. It would be natural to hope to use a spatial model to borrow strength from nearby grid squares. However, the high variability in the quantile correction for large values (occurring since there are relatively few high observations of precipitation) makes it difficult to fit a spatial functional model. Instead we will focus on calibrating directly quantities of higher interest for adaptation, namely high quantiles, where the full calibration did somewhat better.

4 Calibrating individual quantiles

We now focus on calibrating a fixed quantile, q, over a time period of r days. Because of the cross-validation comparison we will perform in the next section, we denote these time periods as folds. An observation $Y_{i,k}^q$ is the qth empirical quantile at location $\boldsymbol{s}_i$ and fold k. That is,

$$Y_{i,k}^q = \widehat{G}_{i,r(k-1)+1:rk}^{-1}(q), \tag{2}$$

where $\widehat{G}_{i,r(k-1)+1:rk}^{-1}(q)$ is the left inverse of the empirical density function made from observations corresponding to the days of fold k, $Y_{r(k-1)+1:rk}$. To link the downscaled precipitation to $Y_{i,k}^q$, we construct

$$X_{i,k}^q = \widehat{F}_{i,r(k-1)+1:rk}^{-1}(q). \tag{3}$$

The goal is now to predict the spatial field $Y_{i,k}^q$ using a calibrated version of $X_{i,k}^q$, where the calibration is estimated using all data that are not in fold k, that is, $Y_{i,-k}^q$ and $X_{i,-k}^q$, where the subscript $-k$ stands for all folds except the kth.

We will use the notation $\boldsymbol{Y}_k^q$ for the vector $[Y_{1,k}^q, \ldots Y_{S,k}^q]$, and similarly for $\boldsymbol{X}_k^q$. Further, we will denote a diagonal matrix with diagonal entries $\boldsymbol{b}$ by $\mathrm{diag}(\boldsymbol{b})$.

As a baseline method, we use the empirical quantile at each location $Y_{i,-k}^q$ as the predictor of $Y_{i,k}^q$. We will later denote this as Model 0, and it should be noted that this prediction does not use the downscaled data. However, it should be a reasonable prediction assuming that the climate is stationary.

As a reference model, we use the smoothing spline method that performed well in Gudmundsson et al. (2012). We will later denote this as Model Ref. The method matches (all) quantiles of the model output to (all) quantiles of the observations using a cubic spline regression, for days with non-zero precipitation.

4.1 Model 1: linear regression

As a first method for doing the calibration, we do linear regression with $X_{i,k}^q$ as covariate. Since the model is for precipitation data, which is asymmetric and positive, we formulate the regression in log scale as

$$\log(Y_{i,k}^q) = \alpha + \log(X_{i,k}^q)\beta_i + \varepsilon_{i,k},$$

where $\varepsilon_{i,k} \sim N(0, \sigma^2)$. Note that we have one parameter β_i for each location, and thus have a spatially varying calibration of the downscaled data. The parameter estimates $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ are estimated by ordinary least squares, and we use $\hat{\boldsymbol{Y}}_k^q = \exp(\hat{\alpha} + \mathrm{diag}(\log(\boldsymbol{X}_k^q))\hat{\boldsymbol{\beta}})$ as a predictor for $\boldsymbol{Y}_k^q$. That is, we use the median as a point estimate (and not the mean). Finally, to apply the method to other data sets as discussed in Sect. 3.2, one simply replaces $\boldsymbol{X}_k^q$ with $\boldsymbol{X}_k^{H,q}$.

4.2 Model 2: incorporating the spatial dependence

There is clearly spatial dependence in the data, which we want to incorporate in the model to improve the predictions. We can do this by assuming that the regression coefficients are spatially dependent, using a stochastic model as follows

$$\log(\boldsymbol{Y}_k^q) = \alpha + \mathrm{diag}(\log(\boldsymbol{X}_k^q))\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k \tag{4}$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}(\nu, \kappa, \phi)), \tag{5}$$

where again $\boldsymbol{\varepsilon}_k \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{\Sigma}_{ij} = C(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$, where C is a Matérn covariance function (Matérn, 1960):

$$C(d) = \frac{\phi^2 2^{1-\nu}}{\Gamma(\nu)}(\kappa d)^\nu K_\nu(\kappa d).$$

Here ϕ determines the variance of the process, ν is a shape parameter of the covariance function, and κ determines the correlation range.

A more computationally efficient alternative to the covariance-based model would be to use a Markov random field prior on $\boldsymbol{\beta}$, similar to that by Bolin et al. (2009). However, this is not needed since the data are measured only at 777 spatial locations, and we therefore use the simpler covariance-based approach here.

The model parameters $\boldsymbol{\theta} = \{\alpha, \kappa, \sigma, \phi, \nu\}$ are estimated using maximum likelihood. The log-likelihood function is given by

$$L(\boldsymbol{\theta}; \boldsymbol{Y}) = \frac{1}{2}\sum_{j\neq k}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}\sum_{j\neq k}(\log(\boldsymbol{Y}_j^q) - \alpha)^T\boldsymbol{\Sigma}_j^{-1}(\log(\boldsymbol{Y}_j^q) - \alpha),$$

where $\boldsymbol{\Sigma}_j = \mathrm{diag}(\log(\boldsymbol{X}_j^q))\boldsymbol{\Sigma}(\nu, \kappa, \phi)\mathrm{diag}(\log(\boldsymbol{X}_j^q)) + \sigma^2\boldsymbol{I}$. We find the ML estimates $\hat{\boldsymbol{\theta}} = \{\hat{\alpha}, \hat{\kappa}, \hat{\sigma}, \hat{\phi}, \hat{\nu}\}$ of the parameters using numerical optimization of the log-likelihood function. Specifically, the function `optim` in R Core Team (2015) is used for the optimization; see the source code at Bolin et al. (2016) for further computational details.

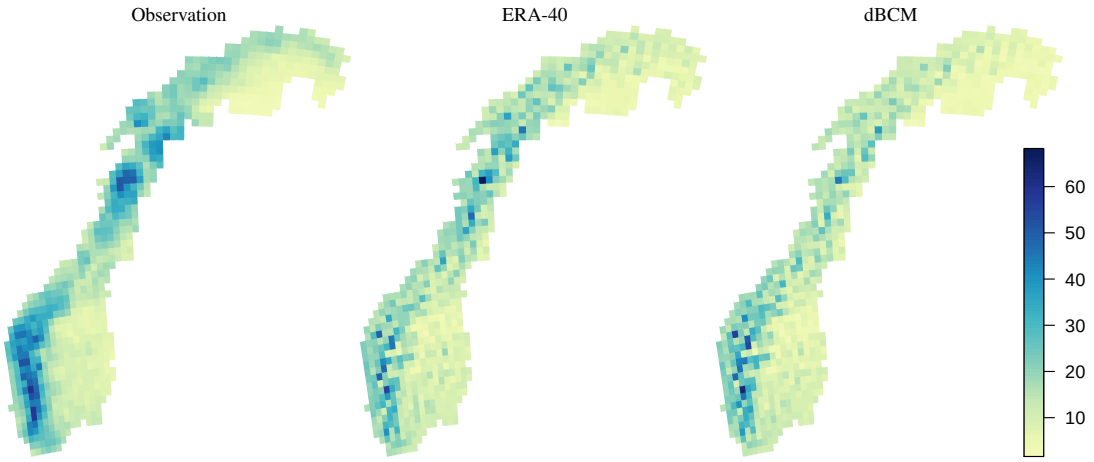

Figure 2. Pointwise 95 % quantiles of OBS (left), dERA40 (middle), and dBCM (right) for the first 5-year period in the cross-validation.

Table 1. Cross-validation mean square error for the 95 % quantile. The best model for each fold is displayed in bold.

k	Model 0	Model 1	Model 2	Model 1s	Model 2s	Model Ref
1	13.29	9.68	8.06	8.59	**7.91**	10.69
2	31.67	7.83	13.17	**6.16**	14.97	6.57
3	14.45	11.23	**8.70**	10.72	9.15	14.11
4	13.54	6.04	7.85	**5.29**	8.61	5.95
5	10.93	5.63	6.59	**4.99**	7.22	5.79
6	23.90	10.15	12.11	**8.37**	13.55	9.78
7	25.47	11.84	14.79	**9.50**	15.83	10.15
8	16.35	10.72	11.47	**9.54**	12.03	10.96
Mean	18.70	9.14	10.34	**7.89**	11.16	9.25

The predictor of $\boldsymbol{Y}_k^q$ is obtained as $\hat{\boldsymbol{Y}}_k^q = \exp(\hat{\alpha} + \mathrm{diag}(\log(\boldsymbol{X}_k^q))\hat{\boldsymbol{\beta}})$, where

$$\hat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{Y}_{-k}^q, \hat{\boldsymbol{\theta}}) = \left(\boldsymbol{\Sigma}(\hat{\nu}, \hat{\kappa}, \hat{\phi}) + \frac{1}{\hat{\sigma}^2}\sum_{j\neq k}\mathrm{diag}(\log(\boldsymbol{X}_j^q)^2)\right)^{-1}$$
$$\sum_{j\neq k}\frac{1}{\hat{\sigma}^2}\mathrm{diag}(\log(\boldsymbol{X}_j^q))(\log(\boldsymbol{Y}_j^q) - \hat{\alpha}).$$

To apply the model to other data one simply replaces $\boldsymbol{X}_k^q$ with $\boldsymbol{X}_k^{H,q}$.

4.3 Model 1s and Model 2s: pre-smoothing the covariates

A somewhat surprising feature of the data are that the quantiles of the observed data, $\boldsymbol{Y}$, are spatially smoother than the downscaled climate model output (see Fig. 2). Because of this, it is natural to add a step in the analysis where covariate is smoothed spatially before it is used in the regression model. This is done using the following model

$$\log(\boldsymbol{X}_k^q) = \log(\widetilde{\boldsymbol{X}}_k^q) + \boldsymbol{\varepsilon}_k$$
$$\log(\widetilde{\boldsymbol{X}}_k^q) \sim N(\mu, \boldsymbol{\Sigma}(\nu_x, \kappa_x, \phi_x)).$$

Here $\log(\widetilde{\boldsymbol{X}}_k^q)$ are independent realizations of a Gaussian–Matérn field and $\boldsymbol{\varepsilon}_k \sim N(0, \sigma_x^2\boldsymbol{I})$. We estimate the parameters using numerical maximization of the log-likelihood

$$L(\mu_x, \sigma_x, \nu_x, \kappa_x, \phi_x; \boldsymbol{X}) =$$
$$\frac{8}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{1}{2}\sum_{k=1}^{8}(\log(\boldsymbol{X}_j^q) - \mu_x)^T\hat{\boldsymbol{\Sigma}}^{-1}(\log(\boldsymbol{X}_j^q) - \mu_x),$$

where $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\nu_x, \kappa_x, \phi_x) + \sigma_x^2\boldsymbol{I}$. Model 1s and Model 2s are then obtained by using

$$E(\log(\widetilde{\boldsymbol{X}}_k^q)|\boldsymbol{X}_k^q, \hat{\mu}_x, \hat{\sigma}_x, \hat{\nu}_x, \hat{\kappa}_x, \hat{\phi}_x) =$$
$$\hat{\mu} + (\boldsymbol{\Sigma}(\hat{\nu}_x, \hat{\kappa}_x, \hat{\phi}_x)^{-1} + \frac{1}{\hat{\sigma}_x^2}\boldsymbol{I})^{-1}(\log(\boldsymbol{X}_k^q) - \hat{\mu})$$

instead of $\log(\boldsymbol{X}_k^q)$ as covariate in Model 1 and Model 2 respectively.

Table 2. Cross-validation mean square error for the 95 % quantile using dBCM predictions. The best model (excluding Model 0) for each fold is displayed in bold. For this comparison, Model 0 can be seen as the target value we want to reach with the dBCM-based predictions.

k	Model 0	Model 1	Model 2	Model 1s	Model 2s	Model Ref
1	13.29	17.35	11.76	17.95	**10.53**	19.80
2	31.67	12.34	14.31	**8.40**	13.53	12.15
3	14.45	60.26	33.90	73.64	**31.18**	64.41
4	13.54	17.03	9.34	19.53	**7.77**	20.24
5	10.93	33.23	13.60	42.82	**10.90**	41.06
6	23.90	57.05	37.95	65.99	**35.37**	64.30
7	25.47	98.60	58.07	118.3	**52.40**	114.17
8	16.35	47.72	29.71	55.38	**27.33**	51.96
Mean	18.70	42.95	26.08	50.25	**23.63**	48.52

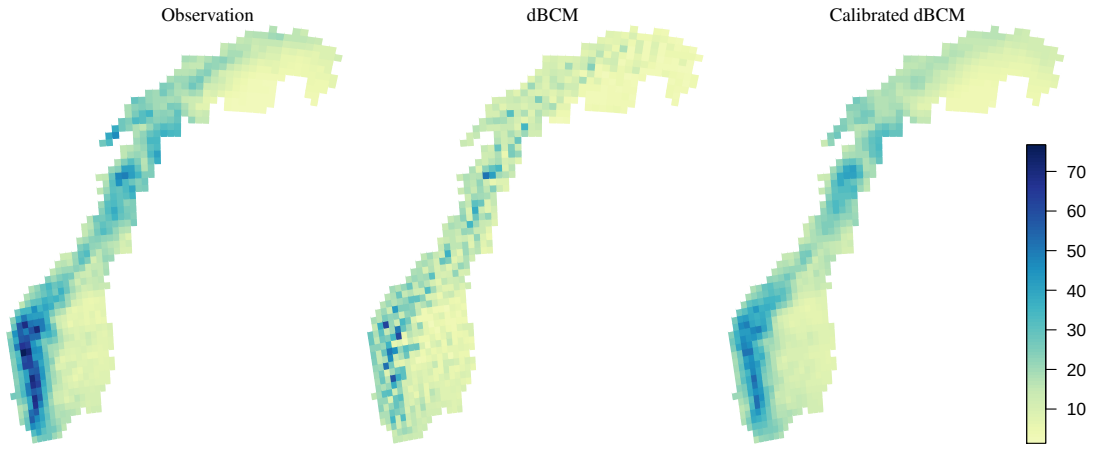

Figure 3. Example calibration for the pointwise 95 % quantiles using Model 2s with the dBCM covariate (right). The result is for the final 5-year time period in the cross-validation study. The observed quantiles (left) and uncalibrated dBCM (middle) are shown as references.

4.4 Results for individual quantiles

In this section, we evaluate the performance of the methods described in Sect. 4 for calibrating individual quantiles. As the tests in Sect. 3 were based on the 95 % quantile, we focus on predicting $Q_{0.95}^{\mathrm{OBS}}$.

The results in Sect. 3 rested on predicting the last 20 % of the data (8 years) based on the first 80 %. Here, we make a more detailed investigation using 8-fold cross-validation to evaluate the performance of the models. The data are divided into eight 5-year periods, and the quantile for each 5-year period is predicted using a model estimated on the rest of the data. The quantiles for the observations, dERA40, and dBCM data for the first 5-year period can be seen in Fig. 2.

The results using the various models can be seen in Table 1, and the results obtained when training the models on the dERA data but using dBCM for prediction are shown in Table 2. One can note that the extension of Model 1 to Model 2 by adding spatial dependency does not improve the results for the dERA40-based predictions, whereas it greatly improves the BCM-based predictions. Furthermore, pre-smoothing improves Model 1 for the dERA40 predictions, whereas it improves Model 2 for the BCM predictions. The reference model Model Ref performs very similarly to Model 1.

Overall, Model 1s performs best for the dERA40 predictions, whereas Model 2s performs best for the dBCM predictions. For the dBCM predictions, Model 2s has satisfactory performance compared with the target performance for that case which is given by the Model 0 results. An example prediction using this model can be seen in Fig. 3.

The results for the different seasons are summarized in Table 3. For all seasons, the conclusion is that Model 1s is preferable if we both train and test the model on dERA40 data, whereas Model 2s is favoured if we train the model on dERA40 data and use that transfer function to calibrate dBCM input. The reason for this is likely that the additional smoothing done in Model 2s compensates for the added uncertainty when using dBCM data in the model trained on dERA40 data.

Table 3. Average mean square error for models trained on dERA40 data. The values in the table are averages across the eight folds for each season.

	Model 0	Model 1s		Model 2s		Model Ref	
RCM	–	dERA40	dBCM	dERA40	dBCM	dERA40	dBCM
Winter	18.70	7.89	50.25	11.16	23.63	9.25	48.52
Spring	8.25	5.98	18.39	6.86	9.54	7.69	21.14
Summer	7.67	5.12	28.80	7.05	9.18	6.99	36.28
Autumn	13.17	9.26	60.08	11.14	20.05	11.92	67.25

5 Discussion

The low quality of Norwegian precipitation in the HIRHAM regional model forced by reanalysis (Orskaug et al., 2011) necessitates a full quantile recalibration/bias correction. Our assessment of such a recalibration on test data indicates that it does a credible job of correcting the dERA40 model, even under changing weather conditions. In order to apply the calibration to climate projections, which is the ultimate goal of this research, we first experiment with the same regional model using a global climate model (GCM), run using historical forcings and corrected using the same calibration as for dERA40. Downscaled global models are unable to describe the observations well. When correcting these downscaled global models, we would of course not expect to get a perfect calibration to data, but would hope that the downscaled GCM/earth system model would describe a similar distribution to that of the observations over a reasonably long period. Unfortunately, this is not the case.

Instead of adjusting the entire distribution, we are able to achieve a better performance by focusing on adjusting an individual quantile. In that case we were able to achieve error rates that indicate that the corrected downscaled climate model performed almost as well as the reanalysis-forced downscaling, indicating that this approach can be a useful tool in downscaling climate projections of precipitation over Norway.

There is a case in between the full quantile adjustment and the individual quantile adjustment, namely simultaneous adjustment of several quantiles. This will be subject to further research.

The sensitivity of regional dynamic downscalings to the lateral boundary conditions is well known (e.g. Rummukainen, 2010, and references therein), and one possibility would be to downscale other reanalyses and compare the results. Since we did not have access to such RCM runs, we were not able to pursue this. On the other hand, we have been able to look at other RCMs (such as the Swedish RCA3 (Samuelsson et al., 2011), with the same reanalysis as boundary condition) and other GCMs (such as the Hadley Centre HadCM3Q0 model). The bias correction based on other regional and global models is very similar to that based on HIRHAM and BCM (results not shown).

6 Data availability

Dynamically downscaled BCM and ERA-40 reanalysis data are accessible from the ENSEMBLES project website http://ensemblesrt3.dmi.dk/ (ENSEMBLES, 2009). Interpolated and gridded precipitation measurements over Norway are available at a $1 \times 1\,\mathrm{km}^2$ spatial resolution from ftp://ftp.met.no/projects/klimagrid/ (METOBS, 2010).

Author contributions. P. Guttorp formulated the Doksum shift calibration method and provided the literature review. E. Orskaug and O. Haug collected and prepared data for the study, and together with I. Scheel and A. Frigessi carried out and documented the analyses in Sect. 3. D. Bolin and J. Wallin were responsible for the individual quantile models of Sect. 4. P. Guttorp combined the description of the two approaches, and all authors contributed to fine-tuning and proofreading the manuscript.

Acknowledgements. We are grateful for helpful discussions with Douglas Maraun, Leibniz Institute of Marine Sciences, Kiel University. This research had partial funding through Statistics for Innovation $(\mathrm{sfi})^2$. Part of this work was done when P. Guttorp visited Chalmers Technical University in 2014, and part when D. Bolin and J. Wallin visited Norwegian Computing Center in 2015.

Edited by: X. Zhang
Reviewed by: two anonymous referees

References

Bjørge, D. and Haugen, J. E.: Simulation of present-day climate in HIRHAM using 'perfect' boundaries, RegClim Techn. Report 1, NILU, 1998.

Bolin, D., Lindström, J., Eklundh, L., and Lindgren, F.: Fast estimation of spatially dependent temporal vegetation trends using Gaussian Markov random fields, Comput. Statist. and Data Anal., 53, 2885–2896, 2009.

Bolin, D., Frigessi, A., Guttorp, P., Haug, O., Orskaug, E., Scheel, I., and Wallin, J.: BiasCorrection code, available at: https://github.com/JonasWallin/BiasCorrection, last access: 16 March 2016..

Christensen, J. H., Boberg, F., Christensen, O. B., and Lucas-Picher, P.: On the need for bias correction of regional climate

change projections of temperature and precipitation, Geophys. Res. Lett., 35, L20709, doi:10.1029/2008GL035694, 2008.

Christensen, J. H., Kjellstrom, E., Giorgi, F., Lenderink, G., and Rummukainen, M.: Assigning relative weights to regional climate models: Exploring the concept, Clim. Res., 44, 179–194, 2010.

Doksum, K.: Empirical probability plots and statistical inference for nonlinear models in the two-sample case, Ann. Statist., 2, 267–277, 1974.

ENSEMBLES project: RT3/RT2B data portal, available at: http://ensemblesrt3.dmi.dk, last access: 1 November 2009.

Fang, G. H., Yang, J., Chen, Y. N., and Zammit, C.: Comparing bias correction methods in downscaling meteorological variables for a hydrologic impact study in an arid area in China, Hydrol. Earth Syst. Sci., 19, 2547–2559, doi:10.5194/hess-19-2547-2015, 2015.

Førland, E., Allerup, P., Dahlström, B., Elomaa, E., Jónsson, T., Madsen, H., Perälä, J., Rissanen, P., Vedin, H., and Vejen, F.: Manual for operational correction of Nordic precipitation data, Tech. Rep. Report 24/96 KLIMA, Norwegian Metorological Institute, 1996.

Furevik, T., Bentsen, M., Drange, H., Kindem, I., Kvamstø, N., and Sorteberg, A.: Description and evaluation of the Bergen climate model: ARPEGE coupled with MICOM, Clim. Dynam., 21, 27–51, 2003.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods, Hydrol. Earth Syst. Sci., 16, 3383–3390, doi:10.5194/hess-16-3383-2012, 2012.

Guttorp, P.: Statistics and climate, Ann. Rev. Statist. Applic., 1, 87–101, 2014.

Jansson, A., Tveito, O. E., Pirinen, P., and Scharling, M.: NORDGRID – a preliminary investigation on the potential for creation of a joint Nordic gridded climate dataset, Tech. Rep. 03/2007, Norwegian Meteorological Institute, 2007.

Kerkhoff, C., Künsch, H. R., and Schär, C.: Assessment of bias assumptions for climate models, J. Climate, 27, 6799–6818, 2014.

Lenderink, G., Buishand, A., and van Deursen, W.: Estimates of future discharges of the river Rhine using two scenario methodologies: direct versus delta approach, Hydrol. Earth Syst. Sci., 11, 1145–1159, doi:10.5194/hess-11-1145-2007, 2007.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M., Venema, V., Chun, K., Goodess, C., Jones, R., Onof, C., Vrac, M., and Thiele-Eich, I.: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, Rev. Geophys., 48, rG3003, doi:10.1029/2009RG000314, 2010.

Matérn, B.: Spatial variation, Meddelanden från statens skogsforskningsinstitut, 49, 1960.

Maule, C., Thejll, P., Christensen, J., Svendsen, S., and Hannaford, J.: Improved confidence in regional climate model simulations of precipitation evaluated using drought statistics from the ENSEMBLES models, Clim. Dynam., 40, 155–173, 2013.

METOBS: Norwegian Meteorological Institute, Climatology Division, met.no gridded climate data $1 \times 1\,\mathrm{km}^2$, version 1.1, available at: ftp://ftp.met.no/projects/klimagrid, last access: 28 June 2010.

Mishra, V., Kumar, D., Ganguly, A. R., Sanjay, J., Mujundar, M., Krishnan, R., and Shah, R. D.: Reliability of regional and global climate models to simulate precipitation extremes over India, J. Geophys. Res.-Atmos., 119, 9301–9323, 2014.

Monjo, R., Chust, G., and Caselles, V.: Probabilistic correction of RCM precipitation in the Basque Country (Northern Spain), Theor. Appl. Climatol., 117, 217–219, 2014.

Orskaug, E., Scheel, I., Frigessi, A., Guttorp, P., Haugen, J. E., Tveito, O. E., and Haug, O.: Evaluation of a dynamic downscaling of precipitation over the Norwegian mainland, Tellus A, 63, 746–756, 2011.

Piani, C., Haerter, J. O., and Coppola, E.: Statistical bias correction for daily precipitation in regional climate models over Europe, Theor. Appl. Climatol., 99, 87–192, 2010.

Räty, O., Räisänen, J., and Ylhäisi, J.: Evaluation of delta change and bias correction methods for future daily precipitation: intermodel cross-validation using ENSEMBLES simulations, Clim. Dynam., 42, 2287–2303, 2014.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: https://www.R-project.org/, last access: 20 June 2015.

Rummukainen, M.: State-of-the-art with regional climate models, Wiley Interdisciplinary Reviews: Climate Change, 1, 82–96, 2010.

Samuelsson, P., Jones, C. G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U., Jansson, C., Kjellström, E., Nikulin, G., and Wyser, K.: The Rossby Centre Regional Climate model RCA3: model description and performance, Tellus A, 63, 4–23, 2011.

Scheel, I. and Hinnerichsen, M.: The Impact of Climate Change on Precipitation-related Insurance Risk: A Study of the Effect of Future Scenarios on Residential Buildings in Norway., The Geneva Papers, 37, 365–376, 2012.

Schmidli, J., Goodess, C. M., Frei, C., Haylock, M. R., Hundecha, Y., Ribalaygua, J., and Schmith, T.: Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps, J. Geophys. Res., 112, D04105, doi:10.1029/2005JD007026, 2007.

Smith, R. L., Berliner, L. M., and Guttorp, P.: Statisticians comment on status of climate change science, AmStat News, 2010.

Teutschben, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, J. Hydrol., 457, 12–29, 2012.

Themeßl, M., Gobiet, A., and Leuprecht, A.: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models, Int. J. Climatol., 31, 1530–1544, 2011.

Uppala, S. M., Kållberg, P. W., and Simmons, A. J.: The ERA-40 reanalysis, Q. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.