Constrained Quantile Regression Splines for Ensemble Postprocessing

JOHN BJØRNAR BREMNES

Norwegian Meteorological Institute, Oslo, Norway

(Manuscript received 6 December 2018, in final form 8 February 2019)

ABSTRACT

Statistical postprocessing of ensemble forecasts is widely applied to make reliable probabilistic weather forecasts. Motivated by the fact that nature imposes few restrictions on the shape of forecast distributions, a flexible quantile regression method based on constrained spline functions (CQRS) is proposed and tested on ECMWF Ensemble Prediction System (ENS) wind speed forecasting data at 125 stations in Norway. First, it is demonstrated that constraining quantile functions to be monotone and bounded is preferable. Second, combining an ensemble quantile with the ensemble mean proved to be a good covariate for the respective quantile. Third, CQRS only needs to be applied to about 10 equidistant quantiles, while those between can be obtained by interpolation. A comparison of CQRS versus a mixture model of truncated and lognormal distributions showed slight overall improvements in quantile score (less than 1%), reliability, and to some extent also sharpness. For strong wind speed forecasts the quantile score was improved by up to 4.5% depending on lead time.

1. Introduction

Ensemble forecasts have now been generated for almost three decades and generally provide a better basis for decision-making processes than single forecasts (Toth and Kalany 1993; Molteni et al. 1996; Richardson 2000). In practice the scenarios are often given a probabilistic interpretation, but model inadequacies tend to limit their direct use. Statistical methods are therefore widely applied to account for deficiencies and make forecasts reliable in a probabilistic sense [see, e.g., Vannitsem et al. (2018) for an overview].

One possible way to classify the methods is whether distributional assumptions on the forecasts are made or not. For the former type the parameters of the distributions are often assumed to be simple functions of the ensemble output and estimated by optimizing the likelihood or the continuous ranked probability score. In recent years the tendency has been toward more flexible distributions or mixture of simpler distributions in order to better describe the relation between ensemble output and what is being observed. For example,

Scheuerer and Hamill (2015) use a censored, shifted gamma distribution for precipitation forecasting, while Baran and Lerch (2016, 2018) use a mixture of truncated normal and lognormal distributions for wind speed forecasting. Mixtures with one component distribution for each ensemble member have also been applied for many years (Raftery et al. 2005).

There is also a variety of methods not assuming any form of the predictive distribution. In analog forecasting, cases similar to the present forecast situation are searched for in long historical archives and applied as forecast (van den Dool 1989; Hamill and Whitaker 2006). Methods based on regression trees have recently also received increasing attention due to their flexibility and ease of use; for example, Taillardat et al. (2016) use quantile regression forest, a slight modification of random forest. A very different strategy is to make linear adjustments of each ensemble member (e.g., van Schaeybroeck and Vannitsem 2015; Wilks 2018). Quantile regression methods, the topic of this article, have been proposed by several authors. Bremnes (2004) introduced local quantile regression which gives noncrossing, nonlinear quantiles. Censored linear quantiles were studied by Friederichs and Hense (2007) for mixed distributed variables like precipitation. Highly adaptable quantile regression functions by means of neural networks were proposed by Cannon (2011). Wahl (2015) used quantile regression in a Bayesian context and considered penalization methods for variable selection. Penalized quantile

Denotes content that is immediately available upon publication as open access.

Corresponding author: John Bjørnar Bremnes, j.b.bremnes@met.no

FIG. 1. (left) Maximum observed wind speed (m s^{-1}) for each of the 125 sites. (middle) Elevation with the darkest gray level representing about 2500 m above sea level. (right) Linear correlation between the ensemble control run (%) and the observation for each site for lead time +36 h.

regression was also applied by Ben Bouallegue (2017). In a related field, Nielsen et al. (2006) used quantile regression with spline functions based on single model NWP output for wind power forecasting.

The main motivation behind this article is that nature imposes few restrictions on the shape of forecast distributions, which can be confirmed by studying the variation in ensemble forecast distributions over time (see section 2). Statistical methods should therefore allow for sufficiently flexible predictive distributions in order to provide well calibrated forecasts in all weather situations. Another motive is that not all information about the future is contained in the ensemble mean and standard deviation, and one aim is to make use of all ensemble members in a versatile manner. To this end, a method based on constrained quantile regression splines (CQRS) is proposed. In CQRS each quantile is represented by a spline function of the ensemble model output, but not necessarily the same output for all quantiles. The latter allows fitted quantiles to better resemble the raw ensemble distribution. Since splines are very adaptable functions, the main challenge is to find appropriate constraints for best possible predictions and in particular for moderate to extreme situations where data are sparse. The method is tested on ensemble wind speed forecasting data at 125 Norwegian sites.

The remainder of this article is structured as follows. The data are presented in section 2. Section 3 describes the constrained quantile regression spline method, two other benchmark methods and the forecast evaluation approach. Section 4 describes experiments and the

outcome of these, while further discussion and conclusions are given in sections 5 and 6.

2. Data

To test the CQRS method, forecasts of wind speed at 10 m height at 125 Norwegian synoptic stations from the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (ECMWF ENS) and corresponding measurements of hourly maximum 10-min average wind speed are considered. The measurement stations were selected from the Norwegian Meteorological Institute's network of automatic synoptic stations with the requirement of at least 97.5% data availability. Most, if not all, of the measurement data have been processed by a simple automatic quality control system and some also manually assessed. For the period under study, the ECMWF ensemble prediction system had a horizontal resolution of about 32 km and consisted of 51 members; one unperturbed (control) and 50 exchangeable, pairwise symmetrically perturbed members. All forecasts were generated at 0000 UTC (coordinated universal time) with lead times +12, +36, +60, +84, and +108 h and bilinearly interpolated to the locations of the stations. The observations and corresponding forecasts are from 21 November 2013 to 31 December 2015. Almost the same data were used in Eide et al. (2017).

The locations of the stations are shown in Fig. 1 along with the topography, the maximum observed wind speed,

and the correlation between the control run for lead time +36 h and the observed wind speed. Norway has a varied topography, with mountain ranges with peaks up to almost 2500 m above the sea level and plateaus, and valleys and fjords as well as lowlands. The strongest wind speeds are observed along the coast and in mountainous areas, while the more sheltered stations in valleys and lowlands are exposed to considerably weaker wind speeds. The skill of the ECMWF ENS forecasts also varies strongly with the locations. In terms of correlation the highest values are seen close to the coast line, while the lowest mostly occur in narrow valleys that are not well resolved in the ECMWF model.

In many statistical methods assumptions on the predictive distributions are made at an early stage in the modeling process. A key aspect is to consider flexibility versus tractability in view of the amount of training data. Despite limited number of ensemble members, studying ensemble forecast distributions over longer time periods gives indications of possible shapes of predictive distributions. Obviously, the ensemble mean and variance/uncertainty vary depending on the weather situations, but further statistics may provide additional information. In Fig. 2 the ensemble skewness, the average of $(\text{member} - \text{mean})^3 / \text{standard deviation}^3$ over the ensemble, is plotted against discretized standardized ensemble means for lead time +60 h. Standardization was obtained by subtracting the temporal mean and dividing by the temporal standard deviation. Clearly, for weak wind cases the majority of forecast distributions are skewed upward, while for strong wind situations the distributions are more often skewed toward zero. Thus, statistical methods should ideally be able to reflect this variability.

For all experiments later the data are split into training and test sets by 1 January 2015, which implies about 13 months of data for training and 12 months for testing and evaluation.

3. Methods

In this section, splines and their properties are first briefly introduced before they are applied in quantile regression. The section then describes ensemble post-processing by means of the truncated normal and log-normal distributions, while the last subsection outlines approaches to forecast verification.

a. Splines

Splines are flexible functions which in their most common form are piecewise polynomials joined together to continuous functions with a certain degree of smoothness. In practice, splines are represented by linear combinations of polynomial basis functions. Due to its many good

FIG. 2. Boxplot of ensemble skewness against standardized ensemble mean for lead time +60 h. There are approximately 1900 forecasts in each of the 50 categories. Outliers are not plotted.

numerical properties B-splines are often chosen. By recursion a normalized B-spline of order $k \geq 1$ is defined as

$$B_j^k(x) = \frac{x - t_j}{t_{j+k-1} - t_j} B_j^{k-1}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+1}} B_{j+1}^{k-1}(x), \quad (1)$$

where $B_j^1(x)$ is

$$B_j^1(x) = \begin{cases} 1 & t_j \leq x < t_{j+1} \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

see, for example, Schumaker (2007, section 4.3). The t_j values are known as knots and control the support and the shape of the function $B_j^k(x)$, while the order k denotes the order (degree + 1) of the polynomials. A spline function $s(x)$ can then be formulated as a linear combination of the B-splines

$$s(x) = \sum_{j=1}^J \alpha_j B_j^k(x), \quad (3)$$

where $\alpha_1, \dots, \alpha_J$ are spline coefficients and $t_1 \leq \dots \leq t_{J+k}$ the underlying knots. From the definition it is possible to derive several useful features that can be taken advantage of. In this work, the two most important are

- if the spline coefficients $\alpha_1, \dots, \alpha_J$ are monotone increasing then the spline $s(x)$ is a monotone increasing function (Schumaker 2007, section 4.9), and
- the spline function $s(x)$ is bounded by its coefficients, that is, $\min\{\alpha_1, \dots, \alpha_J\} \leq s(x) \leq \max\{\alpha_1, \dots, \alpha_J\}$ (Schumaker 2007, section 4.6).

b. Constrained quantile regression splines

Quantile regression was introduced by Koenker and Bassett (1978) and can be seen as a supplement to or even a generalization of regression methods for the

conditional mean. Let $f(x; \alpha)$ be a linear function of one or more covariates x with parameters α and assume it describes how the τ th quantile vary with the covariates x . In quantile regression for the τ th quantile the parameters or regression coefficients α are estimated by minimizing

$$\sum_t \rho_\tau(y_t - f(x_t; \alpha)), \quad (4)$$

where the summation indexed by t is over all training cases and the quantile loss function ρ_τ is given by

$$\rho_\tau(u) = \begin{cases} u(\tau - 1) & u < 0 \\ u\tau & u \geq 0. \end{cases} \quad (5)$$

Note that since the spline $s(x)$ defined by (3) is linear in its coefficients α , the minimization problem is the same when $f(x)$ is replaced by $s(x)$. Koenker and Bassett (1978) showed that the minimization problem could be reformulated as a linear program and efficiently solved using linear programming methods. A very convenient property of linear programs is that additional inequality constraints on the parameters α can be added. Thus, constraints on the spline coefficients as those mentioned above can be imposed.

A potential problem in quantile regression is that quantiles may cross and be invalid since they are fitted separately. Here, this is handled as follows. First, each quantile is assumed to depend only on one covariate which is linearly transformed to the interval $[0, 1]$. Second, the definition of the knots are set equal for all quantiles, which implies that the B-splines are identical. The only difference between the spline representations of the quantiles is then the spline coefficients. Since B-splines by definition are always nonnegative, a spline whose coefficients are higher than another spline, will be the highest of the two. Thus, to avoid crossing quantiles it is sufficient to ensure that corresponding spline coefficients are ordered. Such constraints can be added to the optimization problem by estimating the quantiles in turn. However, brief tests revealed that too many constraints on the coefficients sometimes led to numerical problems. Besides, in the author's experience with local quantile regression (Bremnes 2004) adding similar restrictions may just lead to coinciding quantiles which are not ideal for continuous variables like wind speed. As an alternative, estimation of spline coefficients can be performed without these constraints, but followed by a simple reordering of the coefficients. An additional advantage of the latter is that all quantiles can be estimated in parallel. Note that due to the covariate transformation noncrossing quantiles are only guaranteed on the transformed scale.

FIG. 3. Example of fitted splines at a coastal site for lead time +60 h. For clarity only every 5th quantile is shown. (top) Fitted quantiles on the $[0, 1]$ scale. Blue squares denote the locations of the knots and dashed gray lines the ranges of the training data. (bottom) The same quantiles after transformation back to the original scale. The red lines represent quantile values for two forecast cases. The resulting cumulative distribution functions are shown in the small panel inside with the raw ensemble in gray.

If the same covariate is used for all quantiles, it will also be valid on the original scale. For different covariate definitions we simply resort to reordering of quantiles in the end, if necessary.

An example of fitted quantiles is shown in Fig. 3. For each quantile the corresponding ensemble quantile is used as the covariate and for clarity only 11 out of 51 quantiles are visualized. The upper plot shows fitted quantiles with covariates linearly transformed to $[0, 1]$ and the placements of the knots. Note that the splines are zero beyond the range of the knots, but this does not cause any problem unless new data are far beyond the range of the training data. In any case this can be avoided by careful selection of knots. The lower plot shows the fitted quantile functions transformed back to the original scale and quantiles (red lines) for two randomly

selected forecast days. Their cumulative distribution functions are depicted in the small panel along with the raw ensemble distribution (gray lines).

c. Benchmark methods

By using CQRS, forecast distributions can vary considerably in shape and it is of particular interest to compare CQRS with statistical methods restricting forecasts to certain parametric distributions. As benchmarks, the lognormal EMOS (Baran and Lerch 2015) and the mixture EMOS of truncated normal and lognormal distributions (Baran and Lerch 2016) for wind speed forecasting are chosen. Methods based on other distributions could have been applied as well. Thorarinsdottir and Gneiting (2010) use the truncated normal distribution, while Lerch and Thorarinsdottir (2013) apply the generalized extreme value distribution. Both articles also include intercomparisons of the methods and approaches combining these distributions. Baran and Lerch (2016) conclude that the mixture of truncated normal and lognormal distributions is to be preferred.

Let $\bar{x}$ and s^2 denote the ensemble mean and variance and assume the forecast distribution mean m and variance v are linearly related to the ensemble statistics

$$m = \alpha_0 + \alpha_1^2 \bar{x} \quad (6)$$

$$v = \beta_0^2 + \beta_1^2 s^2 \quad (7)$$

Note that some of the parameters are squared to avoid negative terms in the estimation. For the lognormal distribution (LN) its parameters μ_{LN} and σ_{LN} are related to the distribution mean and variance as follows:

$$\mu_{\text{LN}} = \log\left(\frac{m^2}{\sqrt{v + m^2}}\right), \quad (8)$$

$$\sigma_{\text{LN}} = \sqrt{\log(1 + v/m^2)}. \quad (9)$$

Estimation of the parameters is carried out by minimizing the CRPS using the Nelder–Mead optimization routine (Nelder and Mead 1965) in the statistical programming language R (R Core Team 2018). As initial values in the iterative numerical optimization, linear regression estimates are used for the mean parameters α_0 and α_1 and 0 and 1 for the variance parameters β_0 and β_1 .

For the normal distribution truncated at zero (TN) the location μ_{TN} and scale σ_{TN} parameters are set equal to m and v , respectively. The density function $f(y)$ of the truncated normal and lognormal mixture is given by

$$f(y) = \omega \frac{1/\sigma_{\text{TN}} \phi((y - \mu_{\text{TN}})/\sigma_{\text{TN}})}{\Phi(\mu_{\text{TN}}/\sigma_{\text{TN}})} + \quad (10)$$

$$(1 - \omega) \frac{1}{y\sigma_{\text{LN}}} \phi((\log y - \mu_{\text{LN}})/\sigma_{\text{LN}}), \quad y \geq 0, \quad (11)$$

where ϕ and Φ are the density and cumulative distribution functions of the standard normal distribution. In total there are nine parameters in the mixture which are estimated by maximizing the likelihood using the Nelder–Mead algorithm. The initial values for the mean and variance parameters are obtained as for the LN, while in order to constrain ω to values in $[0, 1]$ it is redefined to $\cos(\omega)^2$ in the optimization (S. Baran 2017, personal communication). Quantiles from the truncated normal and lognormal mixture (TN-LN) distributions are computed by numerical root finding based on the cumulative distribution function.

d. Forecast verification methods

In this article all forecasts are given in terms of a set of quantiles and the overall aim is to assess the quality of these. As with other types of probabilistic forecasts, quantile forecasts should be reliable and at the same time as sharp as possible. Both properties need to be examined in detail separately. In addition, a summarizing measure is essential, for example to be able to rank forecast models. For quantile forecasts the quantile score (QS) is the obvious choice, see Gneiting and Raftery (2007), and references therein. Let y denote the observation and q_τ the forecast for the τ th quantile. The QS for the τ th quantile is then defined as $\rho_\tau(y - q_\tau)$, which is identical to the loss function used in quantile regression, see (4) and (5). For a set of forecasts the QS is just averaged over the cases and, depending on the context, also over the quantile levels. To the latter one should keep in mind that central quantiles have much larger impact on the total quantile score than more extreme quantiles. By integrating over all quantile levels, $\tau \in [0, 1]$, the quantile score equals the continuous ranked probability score. When comparing several forecast models it is advantageous to report scores as skill scores (Murphy 1988). The scores can then be interpreted as improvements relative to the score of a reference forecast system. The quantile skill score (QSS) of a forecast system is defined as $1 - \text{QS}/\text{QS}_{\text{ref}}$, where QS_{ref} denotes the quantile score of the reference system. By definition the skill score of the reference forecast system is zero and positive skill scores indicate improvements over the reference forecast system.

A fundamental requirement for probabilistic forecasts is that they are reliable. By definition the probability of observing a value below the τ quantile is τ .

For a set of quantile forecasts the proportion of measurements below the given quantile is therefore a suitable statistic for reliability. The closer the proportion is to τ the more reliable the forecast model is.

Sharpness refers to the spread of the probability mass and depends only on the forecasts. In most cases sharpness is defined by the length of the interval between two quantiles. Here, sharpness is assessed by calculating the average widths of central intervals with 50% and 88% coverage probabilities. This definition favors unimodal forecast distributions, but in some weather situations the underlying uncertainty may better be represented by multimodal distributions. A supplemental metric for sharpness is therefore also proposed and computed as follows. Assume a forecast in terms of M ordered quantiles is given and that these form $M + 1$ subintervals each with probability mass $1/(M + 1)$. The new metric is then defined by the total length of the m shortest subintervals where m is chosen such that the desired coverage probability is met. For example, to obtain composite intervals with coverages 26/52 (50%) and 46/52 ($\sim 88\%$) m should be set equal to 26 and 46, respectively.

All verification statistics are computed for the whole test period for each station and also averaged over all stations. To further increase the insight of the performance of the forecast models, the data are split into groups by the ensemble mean, ensemble standard deviation, ensemble skewness, and the Hartigan's dip test statistic of unimodality (Hartigan and Hartigan 1985) of the ensemble. Several ways of grouping has been tested, but in section 4 the assessment is restricted to three groups (low, medium, and high) defined by site-dependent 10th and 90th percentiles. Attention is especially paid to reliability, since all forecasts should be reliable and too much averaging may conceal systematic deviations from reliability.

4. Experiments and results

In this section the constrained quantile regression method is tested on the wind speed forecasting data. First, details concerning the splines and the estimation of quantiles are examined. Then, the choice of quantile covariate is considered followed by an intercomparison with the benchmark methods. In the end some experiments for further refinements of the CQRS method are looked into.

a. Spline configurations

Splines are in general highly adaptable functions that would easily overfit the training data unless they somehow are restricted. Several ways of finding the right balance are here tested:

- *Number of interior knots.* The number of knots determines the flexibility of splines. Only equidistant knot sequences with from zero and to three knots in the interval $(0, 1)$ are considered. Brief tests with more knots sometimes resulted in numerical singularities due to sparse data for strong wind speed forecasts.
- *Upper constraint.* Only tests with and without an upper constraint set to 30% higher than the maximum for each station in the training set are carried out. A lower constraint at 0 m s^{-1} is always included. The upper limit was chosen based on a short extreme value analysis on the training data.
- *Increasing constraint.* It seems natural to require quantiles to be nondecreasing with increasing covariate values. With the spline defined in Eq. (3) the requirement is attained if the spline coefficients are in nondecreasing order. In practice there are two ways to achieve this. First, the constraints on the spline coefficients can be added to the linear program optimization. Second, the spline coefficient can be sorted in increasing order after the optimization is done, that is, with no constraints in the optimization step. Both options are tested. Initially, splines without this constraint were also considered, but it did not work out well and was not investigated further.
- *Extrapolation.* Extrapolation refers to predictions at covariate values beyond the range of the training data. Note that extrapolation is not part of the training process, but appears only in the prediction step. Predictions with and without extrapolation is offered. No extrapolation implies that covariates exceeding the training maximum are set equal to the maximum. In practice extrapolation only involves very few cases. Consequently, barely any impact on the summary scores can be expected, so the choice mostly concerns the preferable degree of caution in the forecast models.
- *Spline order.* The order of the B-splines determines how smooth the spline functions are. Linear and cubic splines are here tested.

In total there are 64 combinations. These are all tested by making predictions of the 1/52nd, 2/52nd, ..., 51/52nd quantiles using the corresponding ensemble quantiles as covariate. The models are trained separately for each site and lead time and evaluated on the 1-yr test dataset. The outcome is shown in Table 1 in terms of the quantile score which is averaged over the 51 quantiles. The configurations are ranked by the average rank over the five lead times. As can be seen the differences in quantile scores between the best models are very small, but overall cubic splines with one interior knot and both upper and increasing constraints in the optimization

TABLE 1. Quantile scores ($\times 10^5$) for various spline configurations for lead times +12, +60, and +108 h. Ranks for each lead time are given in parentheses. Scores for only the eight best and the two worst configurations are shown.

Rank	Interior knots	Upper constraints	Extrapolation	Increasing constraints	Spline degree	Lead time (h)		
						+12	+60	+108
1	1	Yes	Yes	Optimization	Cubic	50 789 (1)	56 214 (3)	63 610 (7)
2	2	Yes	No	Optimization	Cubic	50 805 (3)	56 212 (1)	63 656 (15)
3	2	Yes	Yes	Optimization	Cubic	50 804 (2)	56 213 (2)	63 659 (16)
4	1	Yes	No	Optimization	Cubic	50 807 (4)	56 223 (4)	63 612 (9)
5	1	Yes	No	Optimization	Linear	50 912 (16)	56 227 (5)	63 587 (1)
6	2	Yes	No	Optimization	Linear	50 883 (10)	56 236 (7)	63 660 (17)
7	1	No	No	Optimization	Linear	50 913 (17)	56 228 (6)	63 590 (3)
8	2	No	No	Optimization	Linear	50 884 (11)	56 239 (8)	63 662 (18)
63	0	No	No	Order	Cubic	68 050 (61)	76 917 (63)	84 725 (61)
64	0	No	Yes	Order	Cubic	68 084 (62)	76 933 (64)	84 729 (62)

was ranked as number one. In particular it seems that forcing the spline to be nondecreasing in the optimization is important. The performance of the worst spline configurations was about 35% worse than the best. The best model with linear splines with 0 interior knots, roughly linear quantile regression, was ranked 12th (0.2% worse) indicating that nonlinear quantiles are preferred.

The overall best model is for simplicity applied to all sites in the experiments to follow. A sensitivity study was carried out to assess the impact of not using the best individual model for each site. On average the quantile scores were approximately from 1% to 9% worse depending on the lead time. The largest degradation was seen for the shortest lead time.

b. Covariate selection for quantile predictions

In the previous subsection the j th ensemble quantile was applied as covariate for predicting the j th quantile, but this may not necessarily be optimal. Several other covariate definitions are here examined:

- j th ensemble quantile;
- ensemble control member;
- ensemble median;
- ensemble mean;
- the average of the j th ensemble quantile and the ensemble mean;
- the Bernstein polynomial estimator of the j th quantile defined as

$$\sum_{m=1}^M \left[\binom{M-1}{m-1} j^{m-1} (1-j)^{M-m} \right] e_{(m)}, \quad (12)$$

where $e_{(1)}, \dots, e_{(M)}$ are M ordered ensemble members and the estimator can be seen as a weighted average of the ensemble members with the largest weights assigned to members close to the quantile of interest [see Cheng (1995) for further details]; and

- the average of the j th Bernstein quantile estimator and the ensemble mean.

The various definitions are assessed using the best spline configuration from section 4a. The results are summarized in Table 2 in terms of quantile skill score with the model using the ensemble quantile covariate as reference. Using the ensemble control or the ensemble median as covariate gives worse scores than the ensemble quantile covariate, while the ensemble mean overall gives an improvement from 0.11% to 0.47%. Of the quantile based covariates the average of the ensemble quantile and ensemble mean gave the best score, but only slightly better than the two Bernstein polynomial based covariates. Since the former is also faster to compute it was chosen for the rest of the experiments. The spline configuration experiment was also repeated with this covariate definition and it was confirmed that the same configuration as with the ensemble quantile covariate led to the best quantile score.

c. Comparison with other ensemble postprocessing methods

CQRS is in this section compared to the LN model and the TN-LN model described in section 3c. Two CQRS models are fitted; one based on the ensemble quantile covariate (CQRSq) and the other based on the

TABLE 2. Quantile skill scores (%) for a selection of quantile covariates with the ensemble quantile as reference.

	+12	+36	+60	+84	+108
Control	−0.99	−2.13	−3.65	−4.05	−5.57
Median	−0.14	−0.06	−0.15	−0.04	0.03
Mean	0.11	0.32	0.34	0.43	0.47
Quantile + mean	0.15	0.36	0.46	0.52	0.52
Bernstein	0.17	0.38	0.42	0.48	0.45
Bernstein + mean	0.16	0.34	0.44	0.48	0.49

average of ensemble quantile and ensemble mean (CQRSqm). All models were trained separately for each site and lead time and predictions of 51 quantiles were made on the test data. In Table 3 quantile skill scores with TN-LN as reference are given. First, it can be noticed that the raw ensemble forecasts are far worse than all calibrated forecasts. Second, LN was slightly less skillful than TN-LN which confirms the findings of Baran and Lerch (2016). Overall the CQRSqm generated the best forecasts, but only marginally better than TN-LN with the largest improvement (0.7%) at +12 h.

The statistical significance of the latter was tested by means of bootstrap (Efron and Tibshirani 1994) as follows. First, the difference in quantile score between CQRSqm and TN-LN was calculated for each point in time and space (45 477 in total). Brief investigations of dependencies revealed that autocorrelations were more or less insignificant at lag one for the majority of sites, while for some they were clearly significant. For sites close to each other spatial correlation was evident. Based on this, bootstrap sampling with replacement was carried out in blocks of two subsequent points in time and all sites. In total 25000 bootstrap replications were performed. The fractions of these where CQRSqm was better than TN-LN were 100%, 99.9%, 97.7%, 94.8%, and 75.6% for lead times +12 h, +36 h, ... , +108 h. Thus, apart from the longest and maybe second longest lead times, it can be concluded that CQRSqm is better than TN-LN in a statistical sense with respect to the quantile score.

Quantile scores averaged over all quantile levels may hide interesting details. For example, for the statistical models the 50% quantile contributed about seven times more to the overall quantile score than the 98% quantile. That is, the overall quantile score is focused on the central part of the forecast distributions. The quantile skill scores for CQRSqm relative to TN-LN were here larger for the low and high quantiles than on average. For the 98% quantile the QSS varied between 2% and 4% depending on the lead time, while the numbers for the 2% quantile were between 1% and 2%.

Averaging over all forecast situations may also conceal useful information. In Fig. 4 forecasts are grouped according to whether the ensemble mean is low, medium, or high. For both low and high ensemble means the QSS is in clear favor of CQRS with up to nearly 6% improvement, while for the medium group TN-LN is roughly 0.4% better than CQRSqm. A possible explanation could be that the simple linear relation in the location parameters of TN-LN is too restrictive. The LN model is clearly the worst model for low wind speeds, but slightly better than TN-LN for high wind speeds. Grouping the forecasts according to other ensemble

TABLE 3. Quantile skill scores (%) for all models and lead times with the TN-LN as reference. RAW denotes the raw ECMWF ENS forecasts.

	+12	+36	+60	+84	+108
RAW	−67.65	−57.34	−49.07	−41.50	−36.05
LN	−0.36	−0.36	−0.39	−0.30	−0.35
CQRSq	0.56	0.12	−0.22	−0.33	−0.44
CQRSqm	0.71	0.48	0.24	0.19	0.09

statistics gave only minor variations and no obvious patterns. Last, it was investigated whether there were any significant variations in the QSS for CQRS when stations were grouped by maximum observed wind speed and the correlations as in Fig. 1 into low, medium, and high values. Also in this case there were no notable variations.

Reliability was first assessed on an overall basis, but that did not reveal any major differences between the models although the CQRS seemed to be the most reliable. Larger differences were seen when forecasts were grouped by the ensemble mean. In Fig. 5 deviations in quantile reliability (observed proportion minus nominal probability) are shown for lead times +12 and +108 h. Clearly, for low and high ensemble means TN-LN and especially LN has large deviations, while CQRS seems reasonably good. The deviations are largest for the mid quantiles. In general reliability improves with increasing lead time for all models.

Scores for sharpness are shown in Table 4. The ECMWF ENS underestimates uncertainty severely and, consequently, has too narrow forecast intervals. For the 50% coverage interval the LN model has on average the shortest central intervals among the calibrated models. Also the TN-LN produced slightly shorter central intervals than the CQRS methods. For composite intervals, however, the CQRS models are superior. The average length is reduced by roughly 1 m s^{-1} (up to almost 50%) indicating frequent multimodal forecast distributions or clustered quantiles. Since quantiles are estimated separately in CQRS they could perhaps be too clustered by design. The result should therefore be used with caution. For the 88% coverage interval, CQRSqm generated shorter intervals than the other methods both for central and composite intervals. As expected the difference between the two types is less due to fewer configurations to assign the large probability mass. Finally, it can be noticed that the interval lengths somewhat surprisingly did not vary much by lead time. A possible explanation is that the predictability/correlation at some stations is almost independent on lead time. This affects in particular the shortest lead times.

For operational use it is vital that extreme forecasts are physically realistic. By construction the CQRS forecasts

FIG. 4. Quantile skill scores (%) as function of lead time grouped by whether the raw ensemble mean is low (less than the 10th percentile), medium (between the 10th and 90th percentiles), or high (larger than 90th percentile). QSS for the models LN, CQRSq, and CQRSqm are shown based on TN-LN as reference.

are constrained to not exceed wind speeds greater than 130% of the maximum in the training set at each site. For both LN and TN-LN there are no upper limits. As a first check, the number of forecasts with at least one quantile above 45 m s^{-1} , which is approximately the highest wind speed ever recorded in the station network of the Norwegian Meteorological Institute, is computed. For the LN and TN-LN models 52 (0.11%) and 6 (0.01%) forecasts exceeded the 45 m s^{-1} criterion. The highest predicted quantile for TN-LN was about 62 m s^{-1} and similar for the LN model. Next, the number of forecasts with quantiles twice as high as the station maxima are

counted. These were 81 (0.18%) and 4 (0.01%) for the LN and TN-LN models. With a threshold of 1.5 times the observed maxima these increased to 432 (0.95%) and 144 (0.32%), respectively. Although the occurrences are rare they indicate that the upper tails in forecast distributions generated by these methods need further investigations.

d. Refinements

Even though the CQRS approach guarantees non-crossing valid quantiles, separate estimation of many quantiles will in practice often lead to small bumps in the underlying forecast density function usually without

FIG. 5. Deviation from quantile reliability grouped by whether the raw ensemble mean is low (less than the 10th percentile), medium (between the 10th and 90th percentiles), or high (larger than 90th percentile) for lead times (top) +12 h and (bottom) +108 h. The statistic is shown for the models LN, TN-LN, CQRSq, and CQRSqm.

TABLE 4. Average length of forecasts intervals with 50% and 88% coverage probabilities for the models grouped by interval type and lead time. RAW denotes the raw ECMWF ENS forecasts.

	Central interval			Composite intervals		
	+12	+60	+108	+12	+60	+108
50% coverage						
RAW	0.55	1.22	1.82	0.24	0.53	0.78
LN	2.25	2.48	2.75	2.05	2.24	2.46
TN-LN	2.32	2.54	2.82	2.21	2.41	2.66
CQRSq	2.35	2.68	3.01	1.41	1.45	1.50
CQRSqm	2.33	2.62	2.96	1.46	1.57	1.69
88% coverage						
RAW	1.32	2.91	4.32	1.13	2.45	3.62
LN	5.60	6.20	6.92	5.16	5.68	6.28
TN-LN	5.46	6.05	6.77	5.31	5.88	6.52
CQRSq	5.46	6.19	7.09	4.97	5.46	6.08
CQRSqm	5.41	6.01	6.76	4.95	5.40	6.04

any physical reasoning. It is therefore of interest to investigate whether all quantiles need to be estimated. One possibility is to just estimate some quantiles by CQRS and compute the rest simply by interpolation. A brief study where only every 2nd, 5th, 10th, and 25th quantile were estimated by CQRS, while those between were obtained by linear interpolation, was carried out. The lowest and highest quantiles were always estimated by CQRS. Thus, every 2nd quantile, for example, referred to quantiles $1/52$, $3/52$, $5/52$, etc. and similar for the other step sizes. In Table 5 the quantile skill scores with all quantiles estimated by CQRS as reference are shown. Based on these, clearly only every 5th quantile needs to be estimated by CQRS without any loss in forecast skill. A positive side effect is that the computational burden could be reduced by a factor of almost 5 also. Another possibility to achieve smoother underlying distributions is to smooth the regression coefficients, see Koenker (2005, chapter 5) for an overview. This is left for future studies.

The results from section 4a show that almost the same score could be achieved by several spline configurations. To further increase robustness forecasts from these could potentially be combined. Averaging forecasts generated by splines with one and two interior knots gave approximately 0.1% improvement relative to the former. Thus, only minor improvements on the overall performance could be expected. The bootstrap is another averaging technique which is applied in many situations to reduce estimation uncertainty and thereby also increasing the forecast skill. A test with 50 resamples of the training data improved the quantile score of CQRS by only about 0.15% which may indicate that the estimated splines are already quite stable.

TABLE 5. Quantile skill score (%) when estimation of quantiles is partly replaced by linear interpolation. The scores are given for 26, 11, 6, and 3 evenly spaced CQRS estimated quantiles, while the remaining quantiles are obtained by interpolation. The quantile score for CQRS estimation of all 51 quantiles is used as reference.

No. of quantiles	Lead time (h)				
	+12	+36	+60	+84	+108
26	0.013	0.020	0.010	0.010	0.014
11	0.028	0.017	0.032	−0.004	0.046
6	−0.113	−0.172	−0.216	−0.164	−0.148
3	−3.685	−3.456	−3.640	−3.513	−3.674

5. Discussion

In contrast to methods like the TN-LN mixture CQRS does not provide fully specified probability distributions, but only a finite set of quantiles. This limitation may, however, not be of much practical concern. As seen, intermediate quantiles can be obtained by interpolation, while quantiles beyond the range of estimated quantiles could for example be attained by fitting generalized Pareto distributions for the tails (see, e.g., Friederichs 2010). It should also be added that CQRS is not limited to the prediction of the quantiles chosen here. Any set is possible. One only needs to define reasonable covariates. Probabilities for events below or above certain thresholds could be calculated by inverting interpolated quantiles.

There are several studies showing that including several variables from the NWP ensemble may improve the forecasts (see, e.g., Taillardat et al. 2016; Messner et al. 2017; Eide et al. 2017). With the CQRS method it is unfortunately not straightforward to include more variables and further research is required. For two covariates splines in terms of tensor products of B-splines could be used, but further generalization would likely lead to too many parameters to estimate. An alternative would be to take an additive approach with univariate splines, but imposing constraints on the fits may be challenging. It is, however, always possible to reorder quantiles in the end if needed. A third alternative would be to split the process in two stages. First, any regression method for the mean, say, could be applied to make predictions for each ensemble member which at the next step are applied as input to the CQRS.

In future studies it would be interesting to compare CQRS with other quantile regression approaches applied to ensemble weather forecasting, in particular those of Bremnes (2004), Cannon (2011), and Wahl (2015). Concerning computational demand the CQRS likely compares favorably. An R implementation based on the packages *quantreg* and *splines* is computationally very fast. The training of 51 quantiles with nearly

400 training cases took in this study around 0.1 s on a single core on an Intel i7-3770 3.4 GHz CPU (2012 model). Since all quantiles can be estimated in parallel, the time for training could in principle be reduced to a few milliseconds on suitable hardware. Computer codes are available from the author.

6. Conclusions

In this article it is demonstrated how quantile regression with constrained spline functions can be applied to turn NWP ensemble forecasts into well calibrated quantile forecasts. Since assumptions on the shape of the underlying predictive distribution are not made, the CQRS can better adapt to variations in forecast uncertainty than methods based on simple parametric distributions. It also implies that CQRS could be applied to other meteorological variables with at most minor modifications. Further, all ensemble members can be fully utilized which implies that forecast quantiles can resemble the raw ensemble distribution if there is support in the data for this. Another convenient feature is that CQRS is computationally cheap. For future work the main challenge will likely be to find adequate ways of including more than one ensemble variable to the covariates.

The CQRS was here tested on wind speed forecasting data at 125 Norwegian synop stations. On these, results show that CQRS was slightly better, less than 1% in terms of quantile skill score, and more reliable than the truncated normal and lognormal mixture method proposed by Baran and Lerch (2016). For strong wind speed forecasts improvements up to about 4.5% were seen.

Acknowledgments. The author wish to thank Sándor Baran for sharing details on the implementation of the TN-LN mixture method.

REFERENCES

- Baran, S., and S. Lerch, 2015: Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Quart. J. Roy. Meteor. Soc.*, **141**, 2289–2299, <https://doi.org/10.1002/qj.2521>.
- , and —, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130, <https://doi.org/10.1002/env.2380>.
- , and —, 2018: Combining predictive distributions for statistical post-processing of ensemble forecasts. *Int. J. Forecast.*, **34**, 477–496, <https://doi.org/10.1016/j.ijforecast.2018.01.005>.
- Ben Bouallegue, Z., 2017: Statistical postprocessing of ensemble global radiation forecasts with penalized quantile regression. *Meteor. Z.*, **26**, 253–264, <https://doi.org/10.1127/metz/2016/0748>.
- Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, [https://doi.org/10.1175/1520-0493\(2004\)132<0338:PFOPIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2).
- Cannon, A. J., 2011: Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.*, **37**, 1277–1284, <https://doi.org/10.1016/j.cageo.2010.07.005>.
- Cheng, C., 1995: The Bernstein polynomial estimator of a smooth quantile function. *Stat. Probab. Lett.*, **24**, 321–330, [https://doi.org/10.1016/0167-7152\(94\)00190-J](https://doi.org/10.1016/0167-7152(94)00190-J).
- Efron, B., and R. J. Tibshirani, 1994: *An Introduction to the Bootstrap. Monogr. on Statistics and Applied Probability*, No. 57, Chapman and Hall/CRC, 456 pp.
- Eide, S. S., J. B. Bremnes, and I. Steinsland, 2017: Bayesian model averaging for wind speed ensemble forecasts using wind speed and direction. *Wea. Forecasting*, **32**, 2217–2227, <https://doi.org/10.1175/WAF-D-17-0091.1>.
- Friederichs, P., 2010: Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, **13**, 109–132, <https://doi.org/10.1007/s10687-010-0107-5>.
- , and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.*, **135**, 2365–2378, <https://doi.org/10.1175/MWR3403.1>.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, <https://doi.org/10.1175/MWR3237.1>.
- Hartigan, J. A., and P. M. Hartigan, 1985: The dip test of unimodality. *Ann. Stat.*, **13**, 70–84, <https://doi.org/10.1214/aos/1176346577>.
- Koenker, R., 2005: *Quantile Regression. Econometric Society Monogr.*, No. 38, Cambridge University Press, 366 pp.
- , and J. Bassett, 1978: Regression quantiles. *Econometrica*, **46**, 33–50, <https://doi.org/10.2307/1913643>.
- Lerch, S., and T. L. Thorarindottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus*, **65A**, 21 206, <https://doi.org/10.3402/tellusa.v65i0.21206>.
- Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Mon. Wea. Rev.*, **145**, 137–147, <https://doi.org/10.1175/MWR-D-16-0088.1>.
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroligias, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, <https://doi.org/10.1002/qj.49712252905>.
- Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, [https://doi.org/10.1175/1520-0493\(1988\)116<2417:SSBOTM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2).
- Nelder, J. A., and R. Mead, 1965: A simplex method for function minimization. *Comput. J.*, **7**, 308–313, <https://doi.org/10.1093/comjnl/7.4.308>.
- Nielsen, H. A., H. Madsen, and T. S. Nielsen, 2006: Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy*, **9**, 95–108, <https://doi.org/10.1002/we.180>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- R Core Team, 2018: R: A language and environment for statistical computing. R Foundation for Statistical Computing, <https://www.R-project.org/>.

- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, <https://doi.org/10.1002/qj.49712656313>.
- Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- Schumaker, L., 2007: *Spline Functions: Basic Theory*. 3rd ed. Cambridge University Press, 600 pp., <https://doi.org/10.1017/CBO9780511618994>.
- Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>.
- Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, <https://doi.org/10.1111/j.1467-985X.2009.00616.x>.
- Toth, E., and E. Kalany, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, [https://doi.org/10.1175/1520-0477\(1993\)074<2317:EFANTG>2.0.CO;2](https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2).
- van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117**, 2230–2247, [https://doi.org/10.1175/1520-0493\(1989\)117<2230:ANLAWF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2).
- Vannitsem, S., D. Wilks, and J. Messner, Eds., 2018: *Statistical Postprocessing of Ensemble Forecasts*. 1st ed. Elsevier, 362 pp.
- van Schaeybroeck, B., and S. Vannitsem, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, <https://doi.org/10.1002/qj.2397>.
- Wahl, S., 2015: Uncertainty in mesoscale numerical weather prediction: Probabilistic forecasting of precipitation. *Bonner meteorologische abhandlungen*, heft 71, University of Bonn, 108 pp.
- Wilks, D. S., 2018: Enforcing calibration in ensemble postprocessing. *Quart. J. Roy. Meteor. Soc.*, **144**, 76–84, <https://doi.org/10.1002/qj.3185>.