Norwegian Meteorological Institute
met.no

met.no report

Report no. 04/2007
Meteorology
ISSN: 1234-5678
Oslo, February 16, 2007

Improved calibration of precipitation forecasts using ensemble techniques

Part 2: Statistical calibration methods

John Bjørnar Bremnes

Norwegian Meteorological Institute
met.no

report

Number	Subject	Date	Classification	ISSN
04/2007	Meteorology	February 16, 2007	☐ Open ☒ Restricted ☐ Confidential	1234-5678

Title

Improved calibration of precipitation forecasts using ensemble techniques

Part 2: Statistical calibration methods

Authors

John Bjørnar Bremnes

Client(s)	Client reference
EBL Kompetanse	SyPD–2.3_06

Abstract

At sites with measurements or accurate estimates of precipitation, it is often possible to enhance precipitation forecasts by means of statistical methods. The topic of this study is calibration of ensemble precipitation forecasts, and to this end four statistical methods are proposed and tested on real data: (i) transformation of ensemble members such that they in the long run have the same climatology as the observations. (ii) as (i), but preceded by linear regression in order to take into account information about ciculation pattern. (iii) use of scaling factors defined essentially as the ratio of the weighted mean observed precipitation amount and the weighted mean model precipitation amount. (iv) the Bayesian processor of output/ensemble. The first three methods all operate on each ensemble member individually without any regard to other members, while the latter uses all members simultaneously and, thus, has better statistical foundation. The statistical methods are tested at nine sites using ensemble precipitation forecasts by ECMWF's EPS with lead times up to ten days as input. Although the results vary considerably between the sites, the statistical methods generally improves the raw ensemble forecasts considerably – especially for the shorter lead times. Not surprisingly, the Bayesian processor of output/ensemble was on average the best method, both in terms of continuous ranked probability scores and not least reliability/calibration.

Keywords

Statistical calibration, ensemble forecasting, precipitation

Disiplinary signature	Responsible signature
Knut Helge Midtbø, Head Section Meteorology	Øystein Hov, Research director

Postal address
PO Box 43 Blindern
N-0313 Oslo
Norway

Office
Niels Henrik Abels vei 40

Telephone
+47 2296 3000

Telefax
+47 2296 3050

e-mail: met.inst@met.no
Web: met.no

Bank account
7695 05 00601

Swift code
DNBANOKK

1 Introduction

For producers of hydro power accurate precipitation forecasts are of vital importance for smooth production planning and optimal trading. The basis of any precipitation forecast up to about two weeks ahead is mathematical models of the atmosphere, usually called *numerical weather prediction (NWP) models*. These models are based on the fact that if the state of atmosphere is known at given point in time, the laws of physics will foresee the future states. As is well known, weather forecasts, and in particular precipitation forecasts, are not always correct. The causes are essentially twofold; first, the solution of the governing equations of the atmosphere is sometimes highly sensitive to uncertainties in the initial state, which is not possible to determine exactly due to sparse measurements. Second, the NWP models are simplifications of the atmospheric processes. Attempts to quantify the impacts of these causes have led to *ensemble forecasting* in which many forecasts are generated by making small perturbations to the most likely initial state and/or using several parametrization of the physical processes.

Ensemble forecasts provide additional information compared to deterministic forecasts, but when interpreted as probabilistic forecasts, the degree of calibration is often unsatisfactory. For example, the spread of the ensemble can be too low, indicating too strong confidence, or precipitation amounts can be too small or large on average. At many locations ensemble forecasts can be enhanced by using statistical methods which take advantage of historical data comprised of both observations and ensemble forecasts, and use the relation between them to make well calibrated forecasts.

In this report two types of statistical methods are described and tested at sites of interest to the hydro power community. Statistical methods operating on each ensemble member individually are considered in section 2, while section 3 deals with one using complete ensembles. Methods for validating ensemble forecasts are the topic of section 4. The outcome of applying the statistical methods on real data is reported in section 5, followed by some concluding remarks in section 6.

2 Statistical methods for ensemble members

2.1 Local quantile-to-quantile transformation (LQQT)

2.1.1 Brief description

A major source to imperfect ensemble forecasts is biases in the atmospheric models; for example, the mean forecasted precipitation amounts can be less than the mean observed amounts. The aim of this approach is to remove biases quite generally by constructing transformations such that the adjusted forecasts have the same climatology or marginal distribution as the observations. In short, this is obtained by sorting historical forecasts and observations (separately) and estimating the relationship between them. By applying the fitted relation to a new ensemble forecast, that is to each ensemble member and each data point, a new adjusted

ensemble is made. Although the new probabilities or quantiles not necessarily can be trusted, they may validate better than the raw ensemble forecasts, especially if the latter contains biases.

2.1.2 Details

Let F_Y and F_X denote the cumulative distribution functions for observations and forecasts to be calibrated. Then, by standard probability theory the random variable defined by $F_Y^{-1}(F_X(X))$ has the same distribution as the observations (F_Y). However, this relation is only valid for variables with continuous distributions and thus cannot be applied directly to daily precipitation data. In addition, both distribution functions are unknown and must be estimated. To avoid the latter, it is here proposed to estimate the transformation based on the relation between the ordered samples of the two variables and make appropriate adjustments to deal with no precipitation events.

Assume that a training sample of size n is given where $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$ and $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ denote sorted observations and forecasts, respectively. The transformation may then be estimated on basis of the pairs $(y_{(i)}, x_{(i)}), i = 1, ..., n$. Further, let i_0 be the number of pairs where either the observation or the forecast is zero (or below a lower threshold for precipitation). If $m()$ is the estimated transformation, then for a new forecast x_{new} the adjusted forecast y_{new} is given by

$$y_{new}(x_{new}) = \begin{cases} m(x_{new}) & \text{if } x_{new} > 0 \\ y_{(j)} & \text{otherwise} \end{cases} \quad (1)$$

where j is randomly selected from $\{1, 2, ..., i_0\}$.

The estimation of $m()$ can be accomplished using a variety of methods. In this study, local linear regression is chosen and the estimation is performed as follows. In the neighbourhood of the new predictor x_{new}, it is reasonable to assume that the relation can be represented by a simple linear function $m(x) = \alpha_0 + \alpha_1(x - x_{new})$. Its coefficients are then estimated by minimizing the weighted least square loss function

$$\sum_{i=i_0}^{n} (y_{(i)} - m(x_{(i)}))^2 w(x_{(i)}, x_{new}) \quad (2)$$

where $w()$ is a weight function defined such that cases with forecast values close to x_{new} get more weight than those further away. The definition of $m()$ implies that the estimate $m(x_{new})$ is simply the estimate of α_0. In theory, the minimization problem needs to be solved for every new predictor value, but in practice it is usually sufficient to estimate the relation on a fine grid and apply linear interpolation to make predictions in between. Ideally, $m(x_{(i_0)})$ should be zero, but this constraint is not necessarily obeyed by (2). In practice, however, this is rarely a problem, in particular for large training samples.

2.2 Regression and local quantile-to-quantile transformation (REG+LQQT)

A potential disadvantage with the local quantile-to-quantile transformation method (LQQT), is that only one predictor variable can be applied. In reality, however, the transformation could benefit from being dependent on the weather situation at hand. One possibility would be to divide the training sample into different classes according to a classification of the weather patterns and apply LQQT to each category. With many classes and few data in each the estimated transformation would become quite uncertain and in the end result in poor transformed ensemble forecasts.

An alternative solution would be to reduce many predictors to a single new one and then apply LQQT to this. Variable reduction by means of linear regression is here proposed. The new variable to put into LQQT is then predictions from the regression. More flexible regression techniques like neural networks could also be applied.

2.3 Scaling factor (SCL)

2.3.1 Brief description

A simple form for bias correction would be to scale the forecasts. The easiest option of this kind is to sum all historical observations and similarly all historical forecasts and let the ratio of the sums define the scaling factor. As above, it might be desirable to let the scaling factor depend on the weather situation. To achieve this, the use of weighted sums is suggested.

2.3.2 Details

Again, assume that a training set of size n is available, and let $y_1, ..., y_n$ denote observed precipitation amounts and corresponding $\mathbf{x}_1, ..., \mathbf{x}_n$ realizations of the predictor vector. Note that the precipitation amount r_i from the atmospheric model does not need to be among the predictors $\mathbf{x}$. The proposed scaling factor $s(\mathbf{x})$ is then defined as

$$s(\mathbf{x}) \;=\; \left(\alpha \;+\; \sum_{i=1}^{n} w(\mathbf{x},\mathbf{x}_i)\, y_i\right) \;/\; \left(\alpha \;+\; \sum_{i=1}^{n} w(\mathbf{x},\mathbf{x}_i)\, r_i\right) \tag{3}$$

where $w()$ are weights specified such that the historical events most similar to $\mathbf{x}$ get the largest weights and, thereby, also the largest impact on the estimated scaling factor. For interpretability, it is also assumed that the weights add up to one. The tuning parameter α is introduced to make the scaling factor more robust in cases where precipitation is rare or the amounts modest. When the sums are very small, the α tends to dominate and the scale limits to one, that is, no adjustment of the original forecast.

3 Bayesian processor of ensemble/output (BPE)

A major disadvantage with statistical techniques applied separately to each ensemble member, is that there is no guarantee that the resulting probability distributions will be well calibrated. In contrast, statistical methods using the complete ensemble simultaneously should in theory be able to generate well calibrated distributions. In this section, an approach of this kind called the Bayesian processor of ensemble/output (BPE) is described [1] [2]. Since no precipitation events frequently occurs, the modelling is carried out in two steps; first a model for the probability of precipitation, then a separate one for the distribution of precipitation amount given that precipitation will occur.

3.1 Probability of precipitation

3.1.1 Brief description

The basic idea in the BPE method is to transform each variable to standard normal and carry on as if the joint distribution is multivariate normal; more precisely, Bayes rule is applied to decompose the estimation in simpler tasks. For probability of precipitation, this essentially involves estimation of the distributions of the predictor for precipitation events and no precipitation events, respectively.

3.1.2 Details

Let Y be a binary variable representing precipitation occurrence, and assume that it takes the value one when precipitation occurs and zero otherwise. Further, let $f_0(\boldsymbol{x}^*)$ and $f_1(\mathbf{x}^*)$ denote the densities of the predictor vector $\mathbf{x}^*$ when no precipitation and precipitation are observed, respectively. From Bayes rule and the law of total probability it follows that the probability of precipitation π can be formulated by

$$\begin{aligned} \pi \;=\; P(Y=1 \mid \mathbf{x}^*) \quad &= \quad \frac{\pi_c f_1(\mathbf{x}^*)}{(1-\pi_c) f_0(\mathbf{x}^*) + \pi_c f_1(\mathbf{x}^*)} \\ &= \quad \left[1 \;+\; \frac{1-\pi_c}{\pi_c} \frac{f_0(\mathbf{x}^*)}{f_1(\mathbf{x}^*)}\right]^{-1} \end{aligned} \tag{4}$$

where $\pi_c = P(Y=1)$ is the climatological probability of precipitation. In case $f_0(\mathbf{x}^*)$ and $f_1(\mathbf{x}^*)$ are multivariate Gaussian with correlation matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$, equation (4) is

$$\pi \;=\; \left[1 \;+\; \frac{1-\pi_c}{\pi_c} \frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_0|}\, exp\left(-\frac{1}{2}(\mathbf{x}^{*t}\mathbf{\Sigma}_0^{-1}\mathbf{x}^* \;-\; \mathbf{x}^{*t}\mathbf{\Sigma}_1^{-1}\mathbf{x}^*)\right)\right]^{-1} \tag{5}$$

In practice, predictors based on output from atmospheric models are not Gaussian and transformations of the original predictors $\mathbf{x}$ are therefore necessary. A single multivariate transformation of $\mathbf{x}$ to $\mathbf{x}^*$ is complicated to find, but transformation of each variable is feasible by means of the normal quantile transform [1]. Here, we apply the LQQT approach, section

2.1, in which the observations are replaced by quantiles in the standard normal distribution. It is also possible to express (5) in terms of the original predictors $\mathbf{x}$ by means of standard probability transformations, see [1].

3.2 Precipitation amounts

3.2.1 Brief description

The BPE for continuous variates resembles the one for binary predictands described in the previous paragraph. First, the observed precipitation amounts and each predictor variable are separately transformed to the standard normal distribution. Note that only cases with precipitation occurrence are applied, that is, the distributions are conditioned on that there will be precipitation. Further, it is assumed that the distribution of the predictor vector given the observation can be modeled using multivariate linear regression [4]. Bayes rule is then applied to find the distribution of the observation as a function of the predictor. This distribution is also normally distributed, but transformed back to the original units, it can acquire a variety of shapes. In conjunction with the probability of precipitation, the re-transformed distribution forms the forecast.

3.2.2 Details

Let Y^* and $\mathbf{X}^*$ denote random variables for precipitation observations and predictors for days with precipitation; in practice these will be transformed versions of the original variables. Further, assume that the distribution of Y^* and the relation between $\mathbf{X}^*$ and Y^* can be modeled as follows

$$\begin{aligned} Y^* &\sim \mathcal{N}(\mu_0,\, \sigma_0^2) \\ \mathbf{X}^*|Y^*=y &\sim \mathcal{N}(\boldsymbol{\alpha}+\boldsymbol{\beta}y,\, \mathbf{S}) \end{aligned} \tag{6}$$

It can then be shown that the distribution of $Y^*|\mathbf{X}^*=\mathbf{x}$ is normally distributed with expectation and variance

$$\begin{aligned} E[Y^*|\mathbf{X}^*=\mathbf{x}] &= \mu_0 + \boldsymbol{\beta}^t\sigma_0^2(\sigma_0^2\boldsymbol{\beta}\boldsymbol{\beta}^t+\mathbf{S})^{-1}(\mathbf{x}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mu_0) \\ Var[Y^*|\mathbf{X}^*=\mathbf{x}] &= \sigma_0^2 - \boldsymbol{\beta}^t\sigma_0^2(\sigma_0^2\boldsymbol{\beta}\boldsymbol{\beta}^t+\mathbf{S})^{-1}\boldsymbol{\beta}\sigma_0^2, \end{aligned} \tag{7}$$

see for example [5, Ch. 17.2.3]. By minimizing the mean square error, the parameters in the conditional distribution of $\mathbf{X}^*|Y^*=y$, can be formulated as

$$\begin{aligned} \boldsymbol{\beta} &= Cov(Y^*,\mathbf{X}^*) \,/\, \sigma_0^2 \\ \boldsymbol{\alpha} &= E[\mathbf{X}^*] - \boldsymbol{\beta}\mu_0 \\ \mathbf{S} &= Cov(\mathbf{X}^*) - Cov(Y^*,\mathbf{X}^*)\, Cov(Y^*,\mathbf{X}^*)^t \,/\, \sigma_0^2, \end{aligned} \tag{8}$$

see for example [4, Ch. 7.8]. In practice, all parameters are unknown and plug-in estimates are here used to fully determine the distribution of $Y^*|\mathbf{X}^*=\mathbf{x}$.

As mentioned precipitation observations and predictors are in reality not normally distributed and transformations of these must be applied in order to make the modelling described above realistic. Transformations are carried out separately for each variable by means of the normal quantile transform which here is implemented using the LQQT approach described in section 2.1. The conditional distribution in original units can be derived from the distribution of $Y^*|\mathbf{X}^* = \mathbf{x}$.

3.3 Application of BPE forecasts in hydrological models

Deterministic hydrological runoff models, like the HBV model, need amongst other temporal simulations of precipitation and temperature as input. In this framework, calibration of ensembles using separate BPE models for each lead time may pose problems, since the output from BPE in principle are fully specified probability distributions and the temporal dimension is omitted. By simply sampling from each BPE model and randomly linking the samples in time, the inherent temporal dependencies in the raw ensemble forecasts are ignored. To circumvent this dilemma, one may proceed as follows. First, for each lead time, compute as many quantiles from the BPE model as there are ensemble members. The quantiles should be evenly distributed; for example, if the number of ensemble members are N, then the $1/(N+1), 2/(N+1), ..., N/(N+1)$ quantiles could be chosen. Second, for each raw ensemble member, compute its rank at every lead time. Finally, for each ensemble member, use its ranks to choose corresponding BPE quantiles and link these in time. An example is given in table 1.

Lead time	+30	+54	+78	+102	+126	+150	+174	+198	+222
EPS rank	24	6	20	40	9	11	37	26	13
Percentile	47.1%	11.8%	39.2%	78.4%	17.6%	21.6%	72.5%	51.0%	25.5%

Table 1: An example of the ranks of an ensemble member and its corresponding percentiles as a function of lead time. The percentiles are those that should be linked in time from the BPE in order to form a single temporal simulation of precipitation.

4 Validation methods

Probabilistic forecasts provide more information to users than deterministic forecasts, and are, consequently, also more extensive to validate. The attention will be paid to the two basic properties of probabilistic forecasts, *reliability* and *sharpness*, and a measure summarizing the performances. Rather than restricting the focus to validation of probabilities of precipitation above a few predefined thresholds, as is commonly done, the "complete" probability distributions are here examined. The same validation methodology will be used to evaluate both the proposed statistical calibration methods and the raw ensemble forecasts.

4.1 Reliability

Reliability roughly refers to the forecasting method's ability to make probability distributions that can be trusted. In the case of ensemble forecasts, the forecast probability distribution is often interpreted as a set of quantiles, defined such that the probability mass is evenly distributed between the quantiles. For example, for an ensemble of 50 members, 51 bins or intervals are formed, each with a claimed chance of $1/51$ of containing the future observation. Assessment of reliability therefore amounts to checking whether the real probabilities are approximately equal by simply computing the proportion of observations in each bin over many forecasts. For precipitation ensembles, however, a slight modification is needed due to the fact that several ensemble members may have zero precipitation [3]. In cases where zero precipitation is observed and one or more of the ensemble members are zero, the event is therefore randomly assigned to one of the forecast bins with zero precipitation.

Evaluations of reliability are only carried out visually in this report by means of histograms of the proportions (frequently called *verification rank histograms*). These histograms are not only useful to draw conclusions on reliability, but also to reveal possible weaknesses in the forecast system. For example, a U-shaped histogram indicates that the forecast distribution is too narrow on average, while a dome shape is a sign of too sparse forecast distributions. Further, asymmetric histograms are characteristic for biased ensembles.

4.2 Sharpness

Sharpness is a property depending only on the forecasts, not the observations, and concerns the spread of the probability mass. Intuitively, forecast uncertainty should be as low as possible which is obtained by highly peaked probability densities or rapidly increasing cumulative distribution functions. In practice, these characteristics are measured by the lengths of forecast intervals formed by pairs of quantiles. In the experiments later, the average lengths of the 50% and 90% forecast intervals defined by the 25th and 75th percentile, and the 5th and 95th percentile, respectively, are applied to quantify sharpness.

4.3 Summary measure

Although reliability and sharpness more or less describe the quality of probabilistic forecast systems, one is often confronted with the problem of comparing several forecast methods and ranking them. For this task, summarizing measures are helpful, and a suitable score for probabilistic precipitation forecasts is the *continuous ranked probability score* (CRPS). Assuming the forecast is represented by the cumulative distribution function $F(x)$ and y is the observation, the CRPS is defined as

$$CRPS = \int (F(x) - I(x > y))^2 \, dx \tag{9}$$

where $I()$ is the indicator function which equals one if its argument is true and zero otherwise. The lower limit for the CRPS is zero and good forecasting models are characterized by having

scores as close to zero as possible. For many forecasts, the CRPS is computed for each forecast and then averaged.

In order to compute the CRPS for ensembles of size N, the ordered ensemble members are associated with the probabilities $1/(N+1), 2/(N+1), ..., N/(N+1)$ and linear interpolation is applied to evaluate $F(x)$ between the members. Beyond the range of the ensemble, the cumulative distribution function is for simplicity set to zero (below) or one (above). For ensembles of size 50, as in this report, the cut-off has negligible effect on the score.

5 Experiments

5.1 Data

In order to test the statistical methods, nine sites in the southern Norway were chosen by the hydro power companies, and daily precipitation measurements for the years 2004 and 2005 were made available, see table 2. For the same period, ensemble forecasts from ECMWF's ensemble prediction system (EPS) with a horizontal resolution of approximately 80×80 km^2 were extracted and bilinearly interpolated to the sites. These forecasts were initiated at 00 UTC[1] with lead times +30, +54, ..., +222 hour. The ensemble data comprised the meteorological parameters total precipitation, relative vorticity, wind speed, and wind direction; the latter three all at 850 hPa.

The climatology of each site is presented in table 2 in terms of a few statistics. For all sites, there were clearly observed more dry events (less than 0.2 mm/day) than present in the EPS. Since precipitation from the atmospheric models should be interpreted as the average over a grid pixel, it is reasonable that the ratios are larger than 100%, but perhaps not as large as those perceived here. For the percentiles, there were large variations across the sites.

5.2 Description of experiments

The two years of data for each station was divided in a training set (2004) and a set for testing the methods (2005). The specific implementations of the statistical methods were as follows.

5.2.1 Local quantile-to-quantile transformation (LQQT)

The transformation of the raw EPS precipitation was carried out using only the shortest lead time (+30h) and four randomly selected ensemble members; the latter was mainly adopted to avoid too many redundant observations in the training sample. In the local linear least square regression, the weight function $w()$ was defined as

$$w(x_i, x_{new}) = \begin{cases} (1 - (\frac{|x_i - x_{new}|}{d})^3)^3 & \text{if } |x_i - x_{new}| < d \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

[1] At met.no the 00 UTC run is currently available at approximately 09:45 local normal time.

Name	Latitude	Longitude	# dry events	90 percentile	Maximum
Lysebotn	59.06	6.65	242	136	171
Tustervann	65.83	13.91	214	92	94
Vågslid	59.77	7.37	244	70	203
Syrstad	63.02	9.44	227	87	74
Osen	61.25	11.74	186	107	112
Bygdin	61.33	8.80	394	106	96
Nelaug	58.66	8.63	208	100	151
Varaldset	60.67	8.25	427	71	52
Øyestøl	60.80	7.57	363	50	93

Table 2: List of sites, their locations and climates. The latter (last three columns) are specified by the number of dry events, the 90 percentile and the maximum during the two-year period. For each site, the numbers are relative to the climatology of ECMWF EPS and listed as percentages.

where d is a constant controlling the degree of smoothness in the transformation. For each site, d was chosen, without any extensive testing, to be the distance from x_{new} to the 200th nearest data point in the training sample.

5.2.2 Regression and local quantile-to-quantile transformation (REG+LQQT)

For each site, a linear regression with stepwise forward and backward selection of predictors was first performed on the training sample in order to produce adjusted precipitation forecasts for all data. Every predictor mentioned in 5.1 were offered, including second degree polynomials and first order interactions between variables; wind direction was included in terms of four sine and cosine functions. The output from the linear regressions were then transformed using LQQT as described in the previous paragraph.

5.2.3 Scaling factor (SCL)

In the scaling factor method (SCL), the factors were chosen to be functions of the relative vorticity, the wind speed and the wind direction; no attempts were made to include precipitation amounts. The weight function was simply set to be one for the 20% nearest cases and zero for the remaining. By that more robustness was likely gained for extreme situations. The parameter α was set to 0.5 implying that weather situations with modest precipitation amounts on average got scaling factors close to one. The training data for (SCL) was only comprised of the shortest lead time (+30h) in order to get as accurate relations between observed precipitation and weather patterns as possible.

5.2.4 Bayesian processor of ensemble (BPE)

Originally, it was planned to include the four meteorological parameters available for all ensemble members, that is 4×50 predictors, but brief experimentations concluded that it was not feasible. Instead, two new predictors were created by using the adjusted ensemble obtained by the REG+LQQT method. In the model for probability of precipitation only the mean was applied, while for the conditional precipitation amounts the ensemble maximum was also included.

In order to validate the BPE using the same approach as for the other forecast models, predictions of the $1/51, 2/51, ..., 50/51$ quantiles were made. Note that if τ denotes the quantile of interest and π the probability of precipitation, then this implies that the $(\pi - 1 + \tau)/\pi$ quantile must be estimated in the model for precipitation amount given precipitation occurrence. Estimation was of course only carried out if $(\pi - 1 + \tau)/\pi$ was positive; otherwise the quantiles were set to the lower threshold for precipitation (0.1 mm).

5.3 Results

5.3.1 Reliability

The reliability of the various forecasting methods were examined, as explained in 4.1, by means of verification rank histograms. Figure 1 shows these aggregated over the lead times. Those for the raw EPS, clearly demonstrates that too many observations were below all ensemble members and for some sites also above, but to a less degree. To some extent, the former is due to the different scales of the measurements (rain gauges) and the size of the grid boxes (about 80×80 km^2). Too low spread in the ensemble may also be a reason.

The histograms for the scaling method (SCL) have roughly the same characteristics as those of the EPS; thus, the method seemed to have no positive impact on the reliability. This is slightly surprising, but may be partly due to the small precipitation amounts which more or less are left unaltered by the SCL method. The quantile-to-quantile transformation methods, LQQT and REG+LQQT, on the other hand, should handle small amounts well, but still they were not sufficiently reliable. The histograms for these two methods are more symmetric and also U-shaped which is an indication of too little spread in the adjusted ensemble. The reliability of the Bayesian processor of ensemble (BPE) was superior, with only slight deviations at a few stations.

Figure 2 show similar histograms, but now aggregated over the stations. For the raw EPS and the statistical ensemble member methods, the shortest lead times clearly posed most problems. The main reason may be that the EPS is not designed to be properly calibrated for the two or three first days. As expected, the problem decrease with lead time as the forecast distributions get more similar to the climatology and skill deteriorates. For unknown reasons, the BPE was also somewhat unreliable for the first lead time. One possibility could be that the model for probability of precipitation was not accurate enough. Another is that estimation uncertainty is

ignored which generally lead to too sharp distributions as is the case here.

5.3.2 Sharpness

Sharpness is evaluated on basis of the average lengths of the 50% and 90% forecast intervals, figures 3 and 4, respectively. With the reliability results in mind, it is not surprising that the most reliable forecasting method, BPE, on average had quite long intervals for the shortest lead times compared to those of the raw EPS. To some extent this was also evident for the quantile-to-quantile transformation methods. For the scaling method (SCL), however, no marked dependency on the lead time was observed which is likely due to the fact that the estimated scaling factor is constructed on basis of training samples with lead time +30h only. At two sites, Varaldset and Øyestøl, the raw EPS clearly had positive biases, as the statistical methods were able to produce both more reliable and considerably sharper forecasts.

5.3.3 Overall score

The performances of the methods are summarized in figure 5 in terms of the continuous ranked probability scores. The most notable feature is that for most sites the statistical methods were better than the raw EPS. The improvements were largest for the shortest lead times and decreasing with lead time or as predictability worsen. Since biases, as those in the raw EPS, have stronger influence on the CRPS when the uncertainty is low and statistical methods mainly are able to reduce weaknesses due to biases, this is a reasonable observation. Another and somewhat related explanation is that when predictability decreases, there is generally less potential for improving the score. As a curiosity, it could also be mentioned that the absolute CRPSs (not shown) of the raw EPS decreased with lead time at two of the sites. This is just another indication of the strong negative impact biases may have on short lead times.

Among the statistical methods, the BPE had on average the best scores; the results at Syrstad, Osen and Bygdin were especially convincing. Further, it can be noticed that REG+LQQT was clearly better than LQQT at a few sites which suggests that using information about the synoptic circulation pattern was beneficial.

5.3.4 Heavy precipitation events

For hydro power applications heavy precipitation events are of considerable importance. In figures 6, 7 and 8 the forecasts for the three events with the largest observed precipitation amounts at each station are shown. In view of the good CRPSs for the BPE, it is surprising that the method was not able to predict the extreme cases better. In particular, it seems that the predictions of probability of precipitation were quite poor in some instances. At Bygdin and Osen, however, the predictions are relatively good which may explain why BPE is superior to the other methods at these sites.

In figure 6, it can be noted that the cumulative distribution functions of REG+LQQT a few times were “vertical” for large precipitation amounts. Although this improved the CRPS

for these events, this feature should be regarded as a weakness in the implementation of the method that should be possible to avoid.

At Vågslid, figure 6, two of the events were either very badly predicted or examples of errors in the observational data. If the latter is true, these cases are detrimental for the statistical methods and one might anticipate that the overall scores would have been better compared to the raw EPS.

6 Concluding remarks

The study have demonstrated that it is possible to improve raw ensemble forecasts of precipitation by means of statistical methods, but the improvements vary strongly from to site to site. Well calibrated precipitation forecasts are in particular vital for applications where forecasts directly are employed in succeeding physically based mathematical models like hydrological runoff models. Thus, the role of statistical methods is immediate in such cases.

There is certainly scope for further developments of the statistical methods described here. In the implementation of the quantile-to-quantile transformation method, cases with extreme model precipitation may get very strong influence on the transformation of large amounts. In order to make the transformation more robust in these situations, it would be beneficial to apply some sort of resampling. The normal quantile transform in the BPE would also take advantage of this.

On average BPE had the best performance on our data and is also the approach with the greatest potential. However, our use of the method was not as originally planned; instead of using every ensemble member as predictors, it ended up with using only a few statistics of the ensemble as predictors. In principle, all ensemble members could be employed by letting them have common parameters and maybe even by treating them as independent. Further experience is needed.

Future plans include the use of BPE with multi-model ensembles and more generally how to deal with high-dimensional predictive information. Further, experiments have shown that long homogeneous training samples, that is long periods with the same atmospheric model, substantially can improve the quality of operational weather forecasts [6]. For hydro power applications, it would be interesting to quantify the improvements of precipitation forecasts as a function of the length of training period and relate these to the quality of raw operational forecasts.

Acknowledgment

This work is partially funded through EBL Kompetanse (www.ebl.no).

References

[1] R. Krzysztofowicz and C. J. Maranzano (2006). *Bayesian processor of output for probability of precipitation occurrence.*

[2] R. Krzysztofowicz and C. J. Maranzano (2006). *Bayesian processor of output for probabilistic quantitative precipitation forecast.*

[3] J. McLean Sloughter, A. E. Raftery, and T. Gneiting (2006). *Probabilistic quantitative precipitation forecasting using Bayesian model averaging*. Monthly Weather Review, in press.

[4] R. A. Johnson and D. W. Wichern (1992). *Applied multivariate statistical analysis*. 3rd Edition. Prentice Hall International.

[5] M. West and J. Harrison (1997). Bayesian forecasting and dynamic models. 2nd edition. Springer Verlag.

[6] T. M. Hamill, J. S. Jeffrey and S. L. Mullen (2006). *Reforecasts: An Important Dataset for Improving Weather Predictions*. Bulletin of the American Meteorological Society, Volume 87, Issue 1, pp. 33–46.

Appendix A: Software

The statistical methods described in this report have been implemented using the statistical programming language R and included in an R-package called SWEAP which can be obtained from the author.

Some computer code examples are given below:

```
# load library
library(SWEAP)

# load ensemble data
# use ?lysebotn to get more information
data(lysebotn)


##
##   Calibration by means of quantile-to-quantile transformation
##

# fit (using only the first ensemble member)
k   <- lysebotn$NO == 1
fit <- lqqt.fit(lysebotn$RR.O[k], lysebotn$RR[k], lower=0.1,
                nlls=20, tricube=TRUE)

# plot fitted transformation
plot(fit$x, fit$y, type="l", las=1, xlim=c(0,100), ylim=c(0,100),
     xlab="Forecast (mm)", ylab="Observation (mm)", main="Transformation")
abline(0, 1, lty="dashed")

# calibrate a single ensemble forecast (generated 2005-09-12)
k   <- lysebotn$ATIME == 2005091200
```

```
cal <- lqqt.predict(fit, newx=lysebotn$RR[k])

# plot raw and calibrated forecasts as cumulative distribution functions
# original forecast (red), calibrated forecast (blue) and observation (black)
plot(sort(cal), 1:50/51, xlab="Precipitation", ylab="Cumulative probability",
     type="b", col="blue", las=1)
lines(sort(lysebotn$RR[k]), 1:50/51, type="b", col="red")
abline(v=lysebotn$RR.O[k][1], col="black")
grid()


##
##   Calibration by means of scaling
##

# estimate scaling factor for a single ensemble forecast
ktrain <- lysebotn$ATIME < 2005010100 & lysebotn$NO == 1
ktest  <- lysebotn$ATIME == 2005091200
scl    <- locscale.fit("RR.O", "RR", c("FF.L850","DD.L850","VO.L850"),
                       data=lysebotn[ktrain,], newdata=lysebotn[ktest,],
                       period=c(0,360,0))

# histogram of scaling factors
hist(scl, main="Scaling Factors for Ensemble Members")

# plot raw and calibrated forecasts as cumulative distribution functions
# original forecast (red), calibrated forecast (green) and observation (black)
plot(sort(scl*lysebotn$RR[ktest]), 1:50/51, type="b", col="seagreen3", las=1,
     xlim=c(0,120), xlab="Precipitation (mm)", ylab="Cumulative probability")
lines(sort(lysebotn$RR[ktest]), 1:50/51, type="b", col="red")
abline(v=lysebotn$RR.O[ktest][1], col="black")
grid()


##
##    Calibration by means of Bayesian Processor of Ensemble/Output
##

# load ensemble data
data(lysebotn2)

# compute ensemble mean and max
lysebotn2 <- cbind(lysebotn2,
                   EPS_0_MEAN=rowMeans(lysebotn2[,2:51]),
                   EPS_0_MAX=apply(lysebotn2[,2:51], 1, max))

# probability of precipitation
ktrain   <- lysebotn2$TIME < 2005010100
ktest    <- lysebotn2$TIME == 2005091406
pop.fit  <- bpe2.fit(y=lysebotn2$RR.O[ktrain]>0.1,
                     x=lysebotn2[ktrain,"EPS_0_MEAN",drop=FALSE], sep="_")
pop      <- bpe2.predict(pop.fit, newx=lysebotn2[ktest,"EPS_0_MEAN",drop=FALSE])

# precipitation amounts
k        <- ktrain & lysebotn2$RR.O > 0.1
vars     <- c("EPS_0_MEAN", "EPS_0_MAX")
prec.fit <- bpe.fit(y=lysebotn2$RR.O[k], x=lysebotn2[k,vars], lower.y=0.1, sep="_")

pr       <- (1 - (50:1/51)) / pop    # conditional quantile probabilities
precip   <- bpe.predict(fit=prec.fit, newx=lysebotn2[ktest,vars], qt.prob=pr)

# plot raw and calibrated forecasts as cumulative distribution functions
# original forecast (red), calibrated forecast (green) and observation (black)
plot(sort(precip), 1:50/51, type="b", col="magenta4", las=1,
     xlim=c(0,120), xlab="Precipitation (mm)", ylab="Cumulative probability")
lines(sort(unlist(lysebotn2[ktest,2:51])), 1:50/51, type="b", col="red")
```

```
abline(v=lysebotn2$RR.O[ktest], col="black")
grid()
```

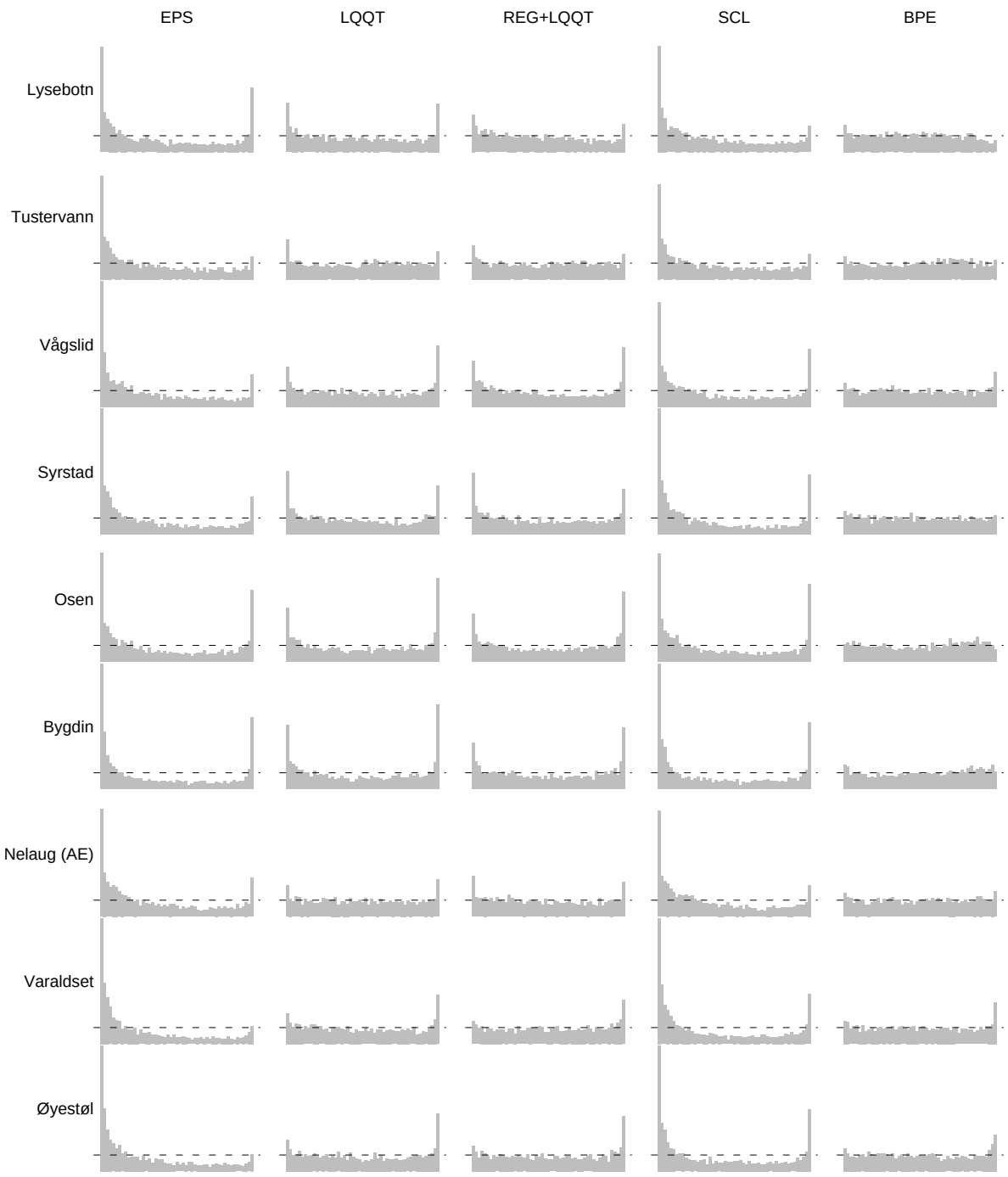

Figure 1: Verification rank histograms for each method and station. The ranks are averaged over the lead times. Well calibrated forecast methods have ranks close to the dashed line. Note that some bars are clipped and should in fact be considerably longer than they appear.

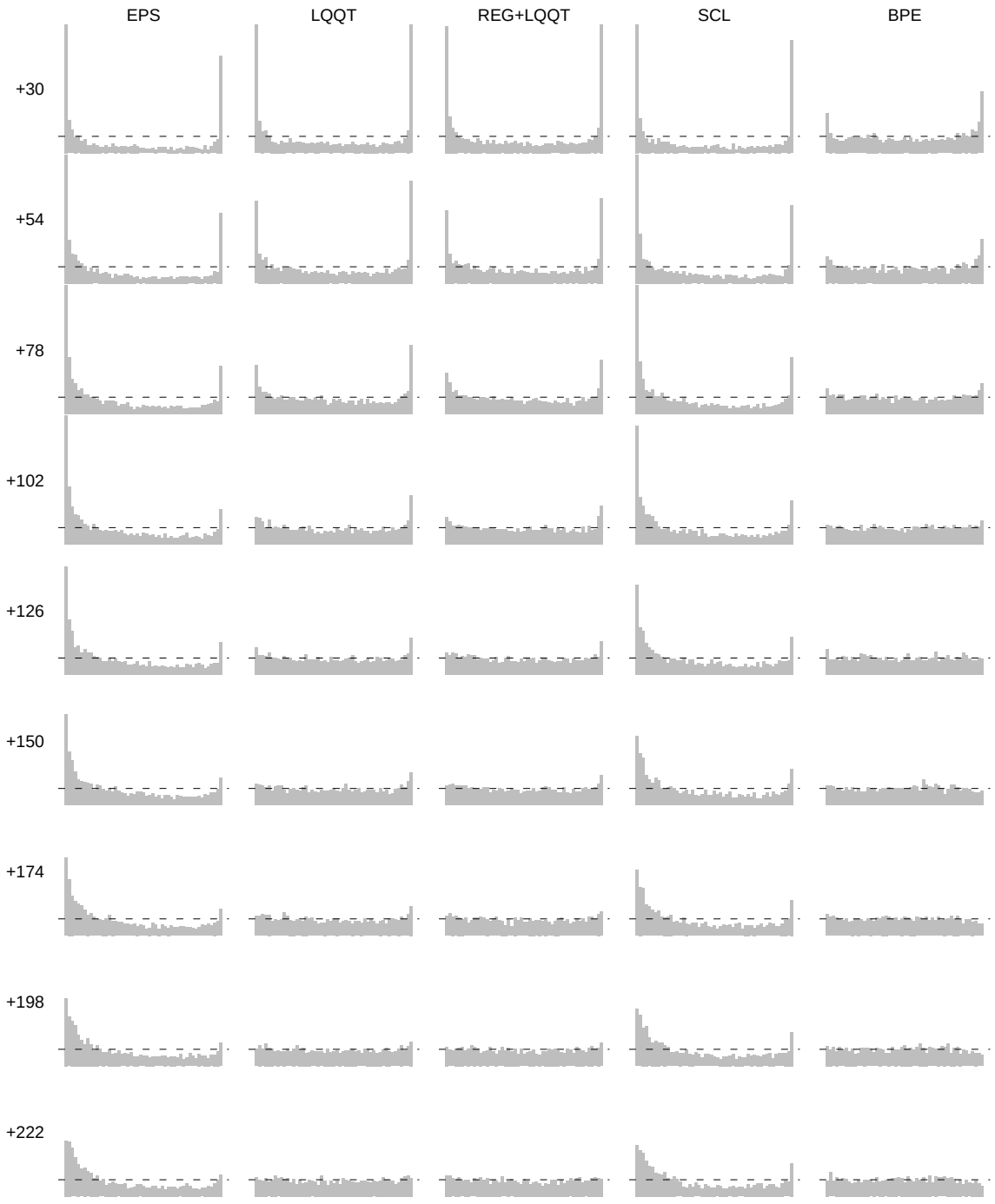

Figure 2: Verification rank histograms for each method and lead time. The ranks are averaged over the nine stations. Well calibrated forecast methods have ranks close to the dashed line. Note that some bars are clipped and should in fact be considerably longer than they appear.

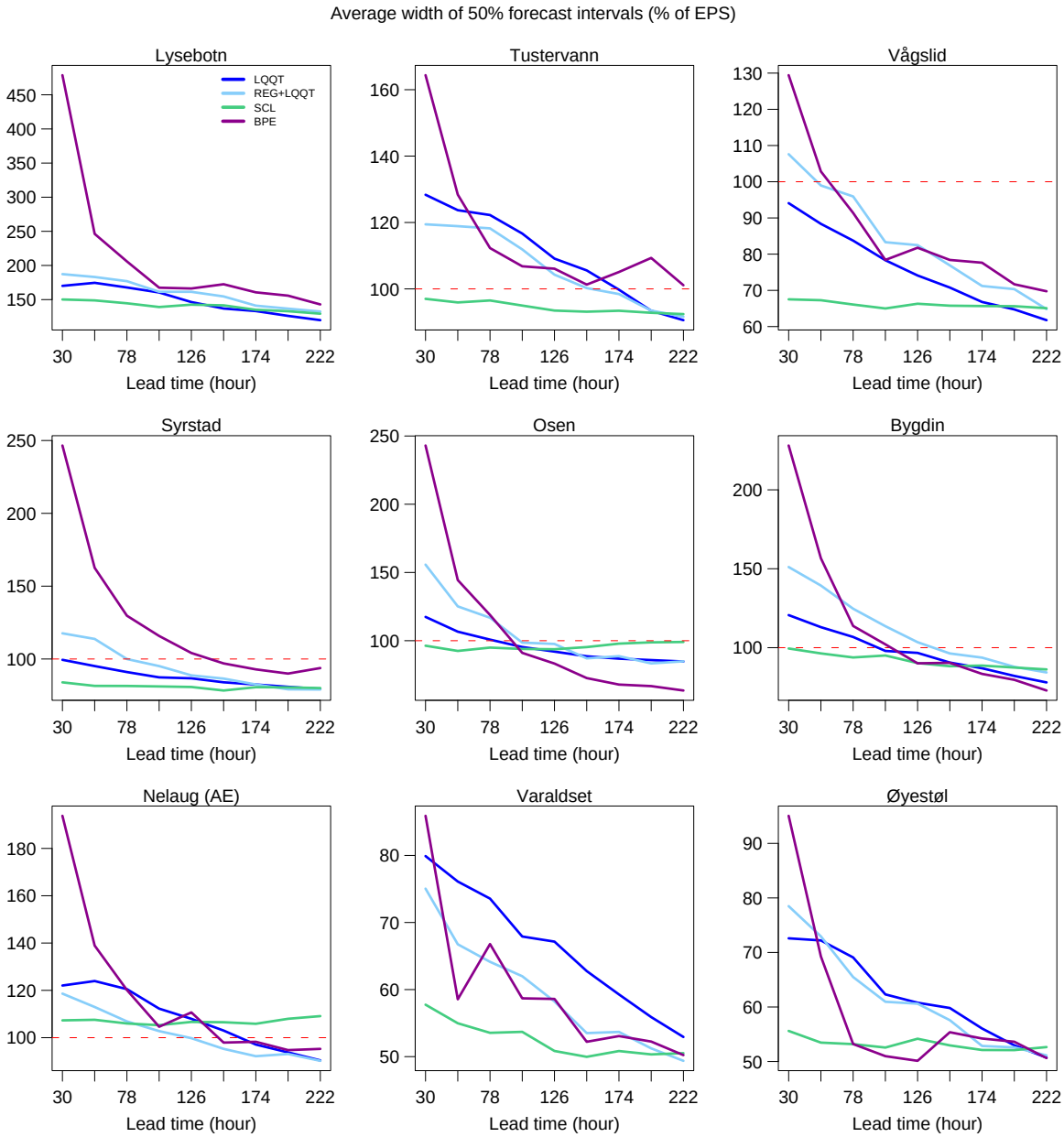

Figure 3: Average lengths of 50% forecast intervals as a function of lead time for the methods LQQT (blue), REG+LQQT (light blue), SCL (green) and BPE (magenta). The interval lengths are specified in percentage of the EPS lengths (red and dashed).

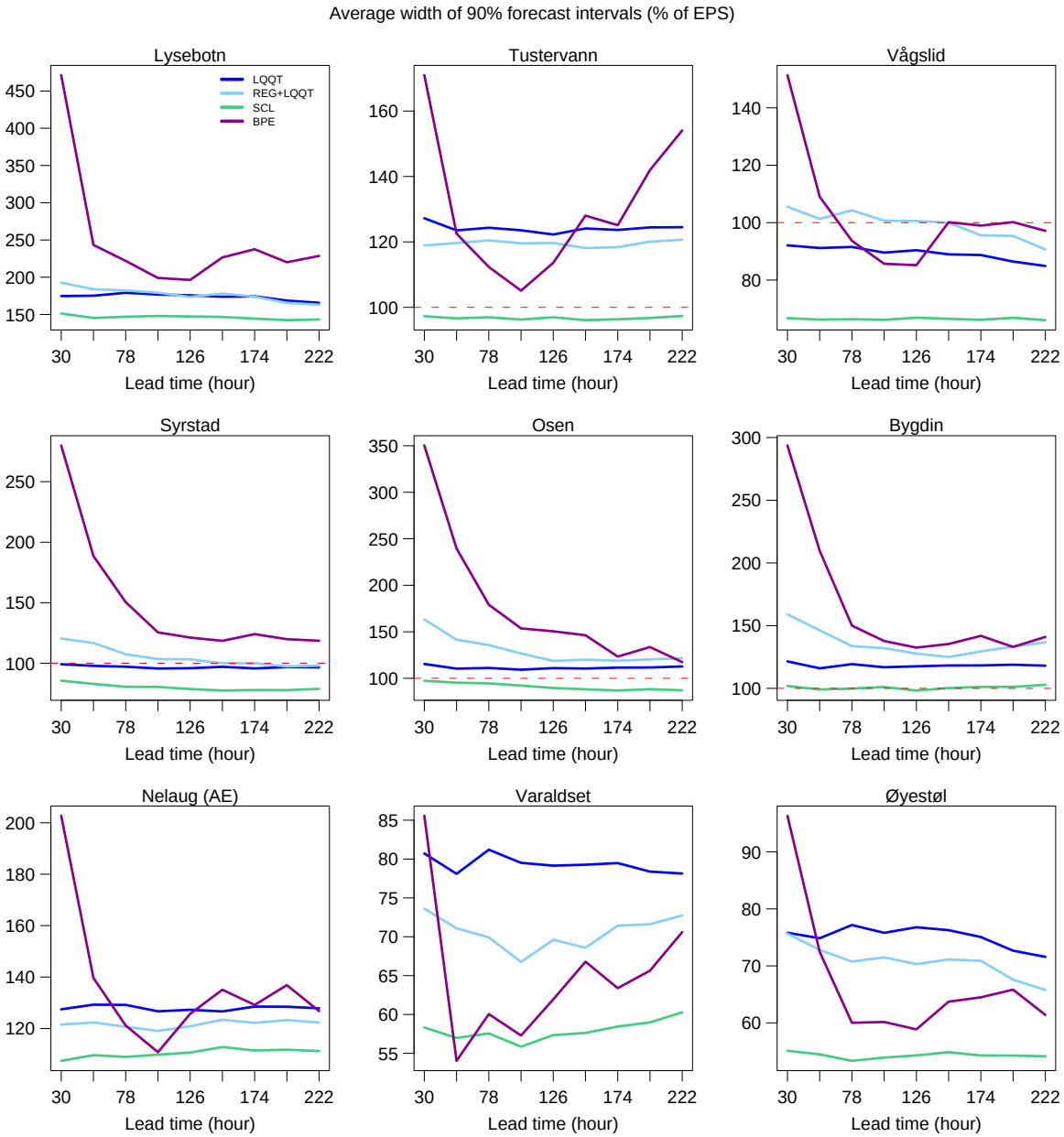

Figure 4: Average widths of 90% forecast intervals as a function of lead time for the methods LQQT (blue), REG+LQQT (light blue), SCL (green) and BPE (magenta). The interval lengths are specified in percentage of the EPS lengths (red and dashed).

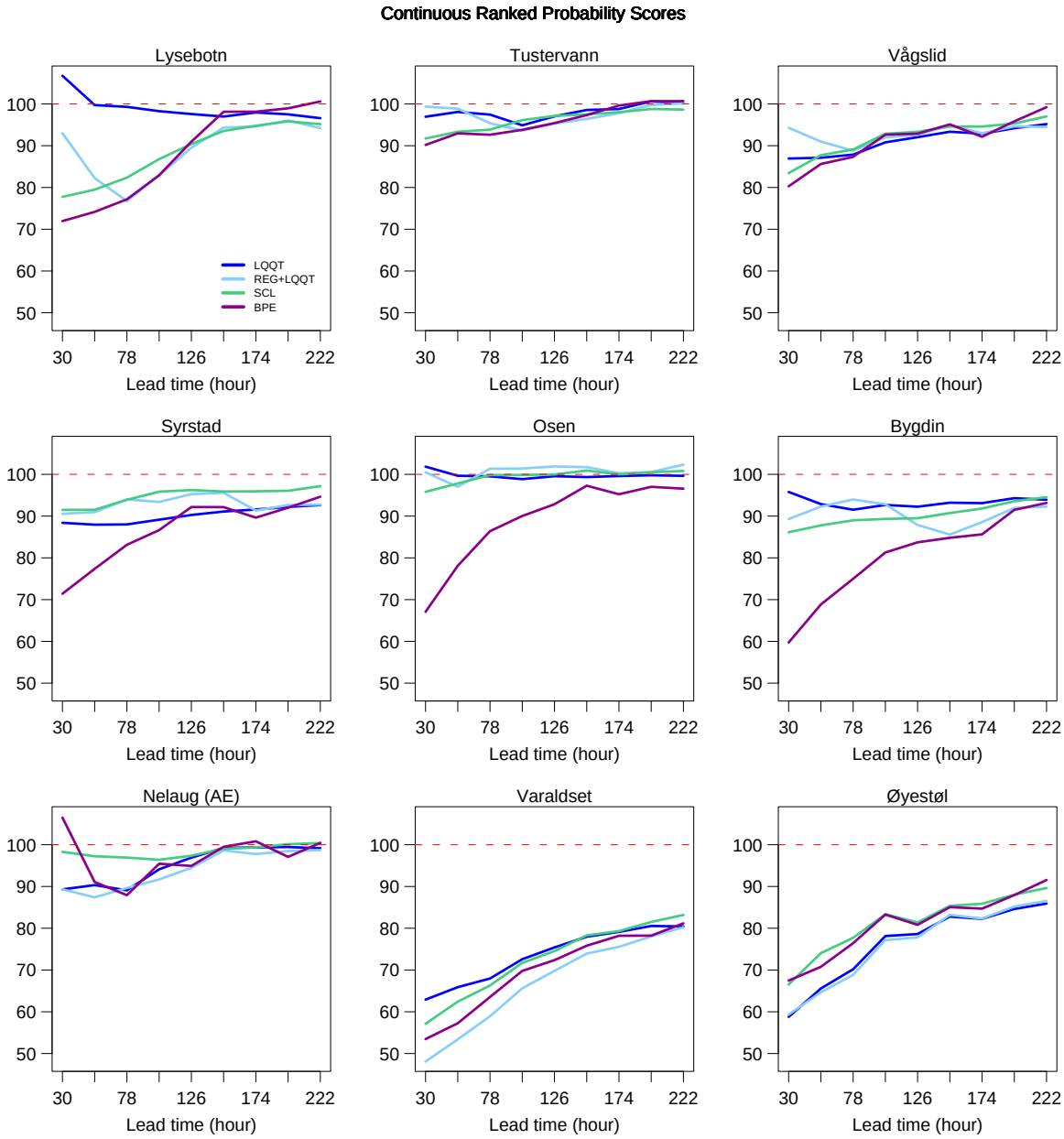

Figure 5: Continuous ranked probability scores as a function of lead time for the methods LQQT (blue), REG+LQQT (light blue), SCL (green) and BPE (magenta). The scores are relative to the scores of EPS (red and dashed). Low scores are best.

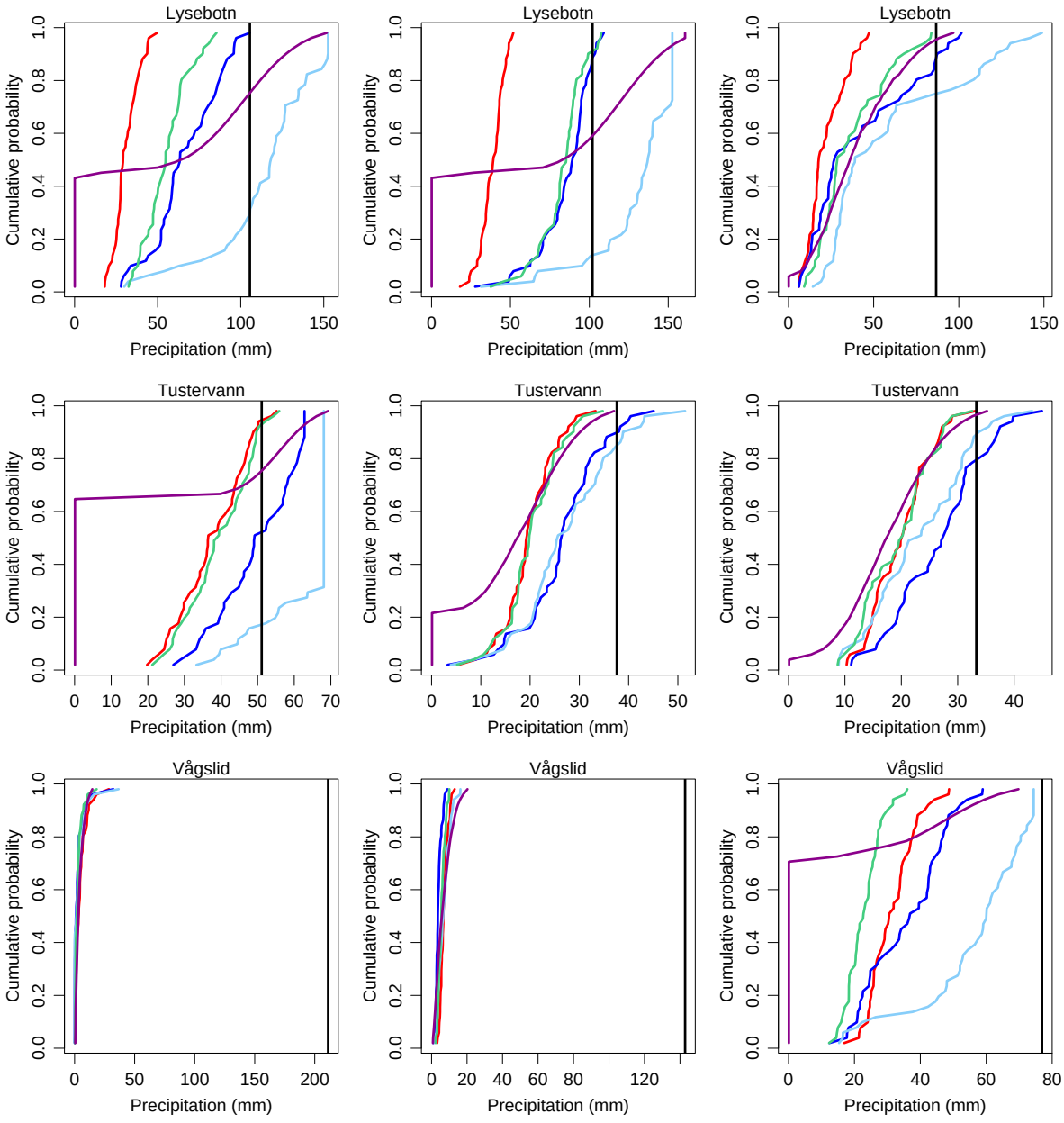

Figure 6: Forecasts (+54h) in terms of cumulative distribution functions for the three cases with largest observed precipitation amounts in the test data. The methods are raw EPS (red), LQQT (blue), REG+LQQT (light blue), SCL (green) and BPE (magenta). The observations are in black.

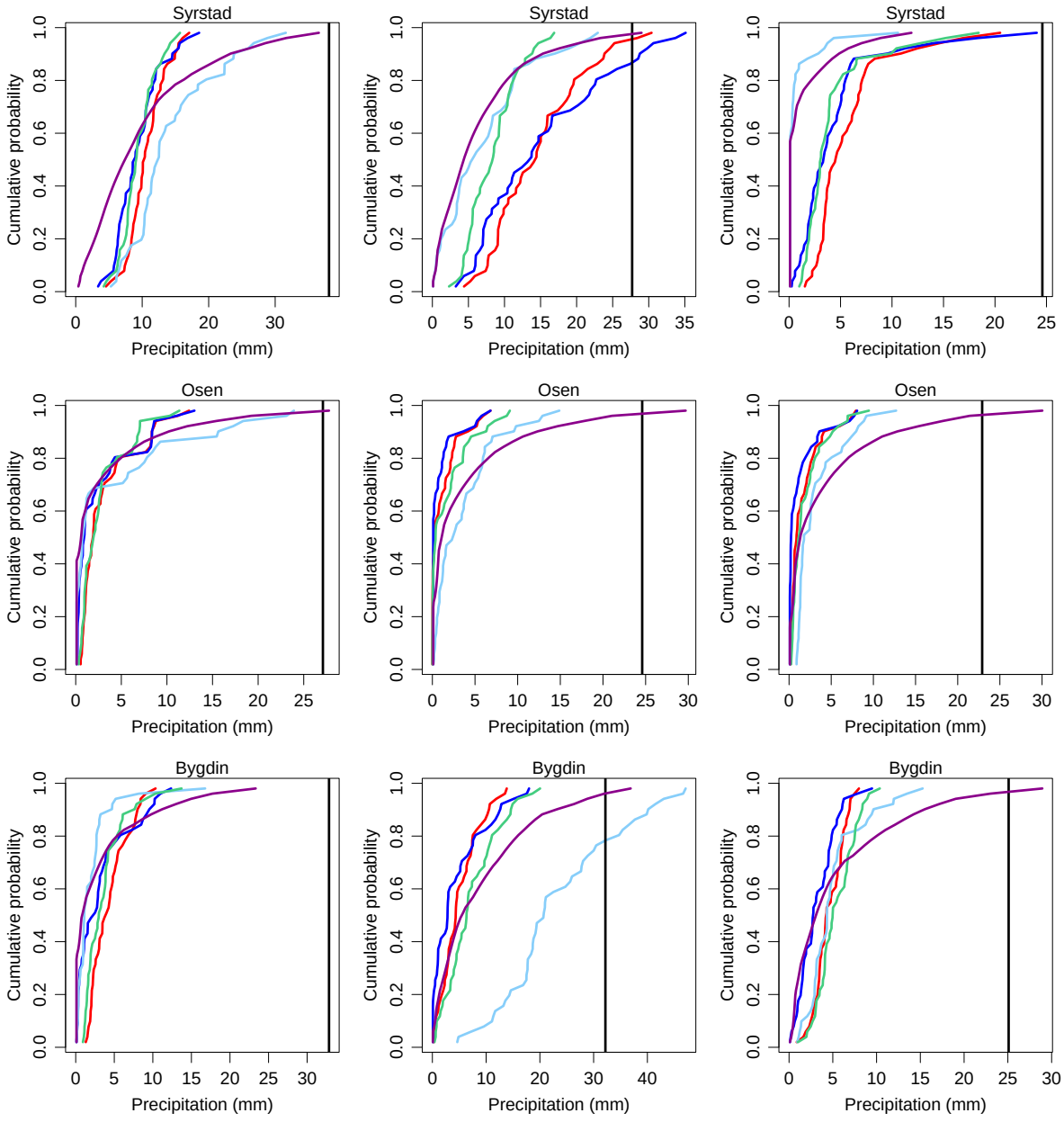

Figure 7: Forecasts (+54h) in terms of cumulative distribution functions for the three cases with largest observed precipitation amounts in the test data. The methods are raw EPS (red), LQQT (blue), REG+LQQT (light blue), SCL (green) and BPE (magenta). The observations are in black.

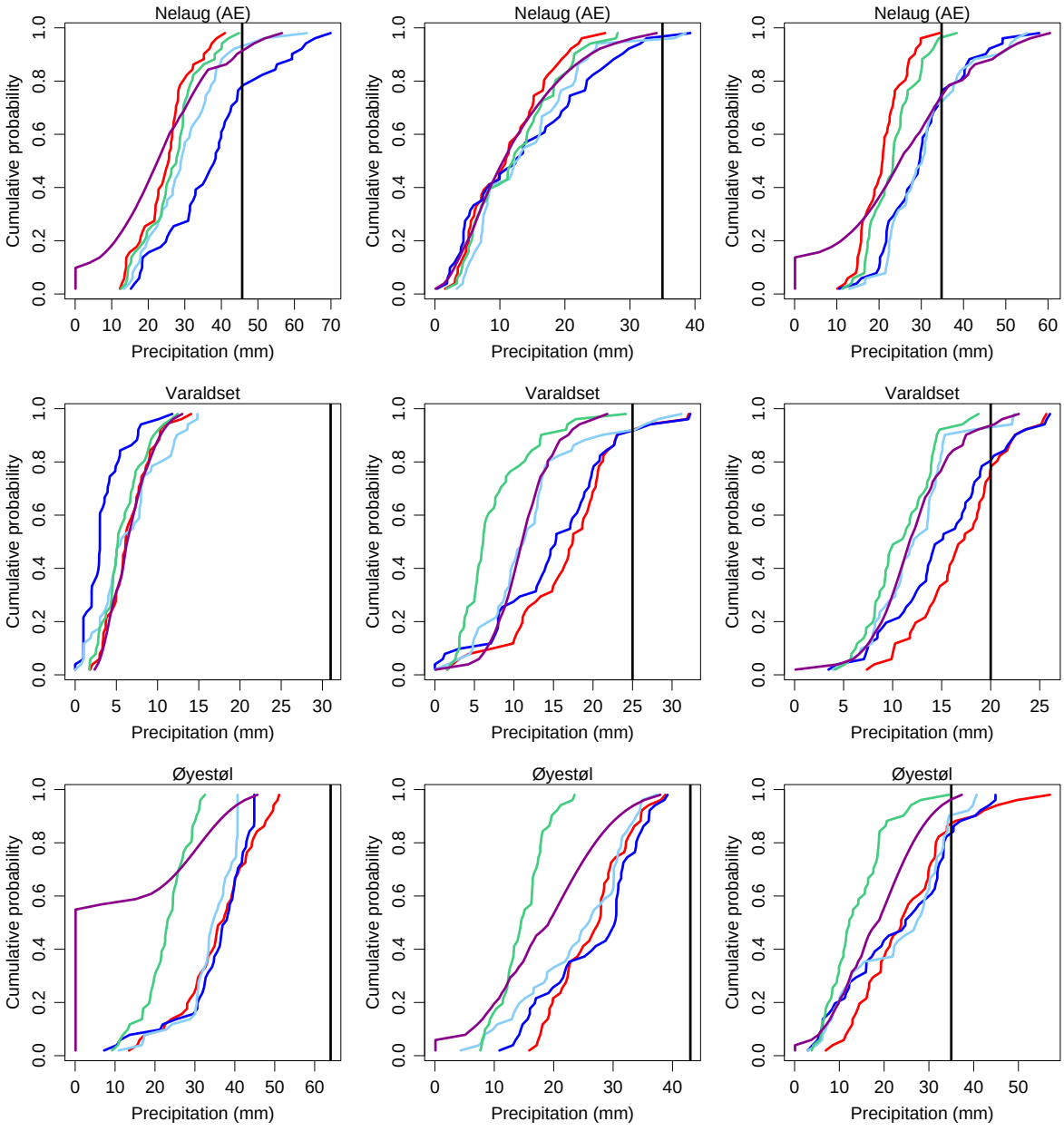

Figure 8: Forecasts (+54h) in terms of cumulative distribution functions for the three cases with largest observed precipitation amounts in the test data. The methods are raw EPS (red), LQQT (blue), REG+LQQT (light blue), SCL (green) and BPE (magenta). The observations are in black.

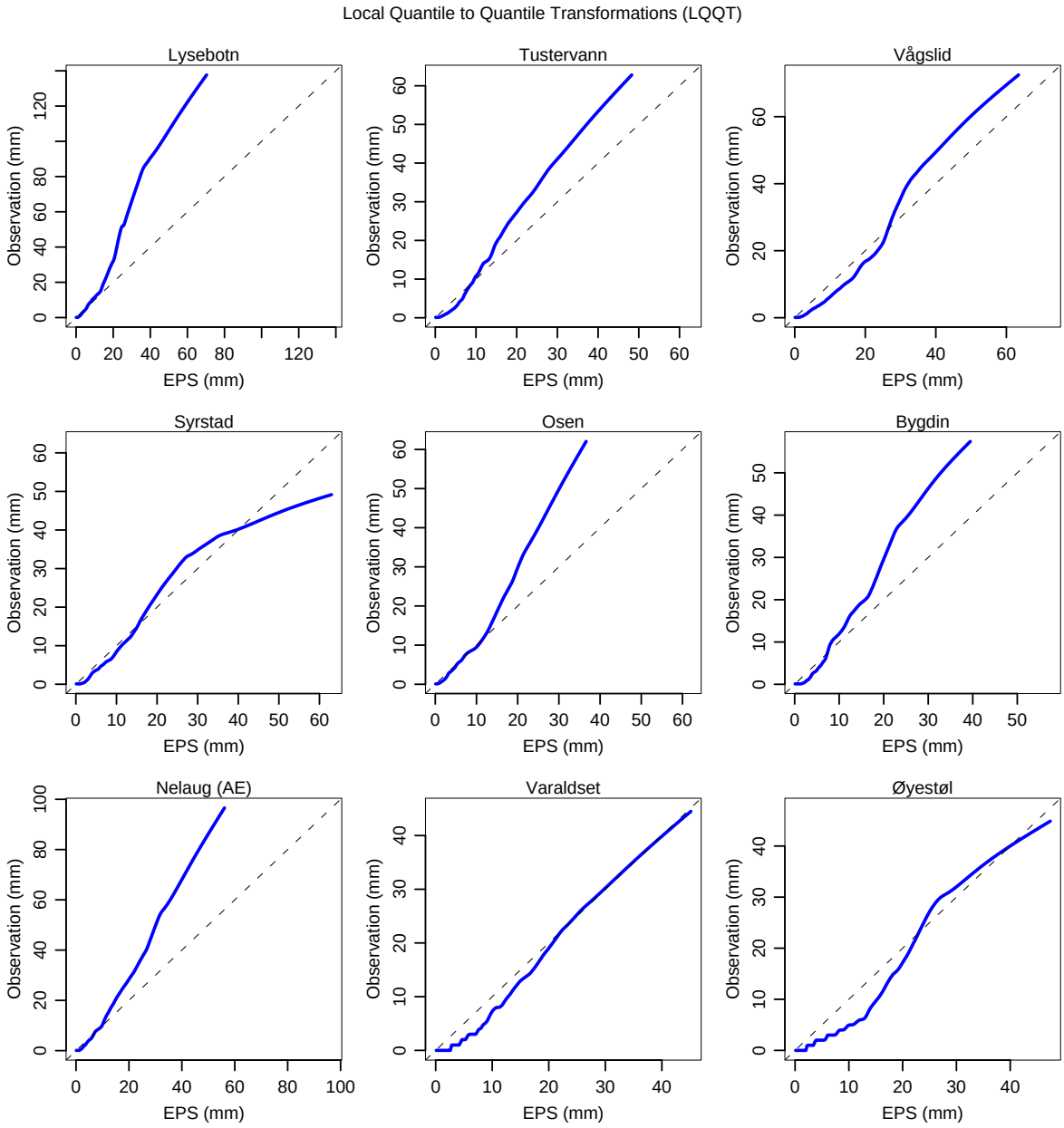

Figure 9: Estimated transformations obtained by the LQQT method for each of the nine sites.

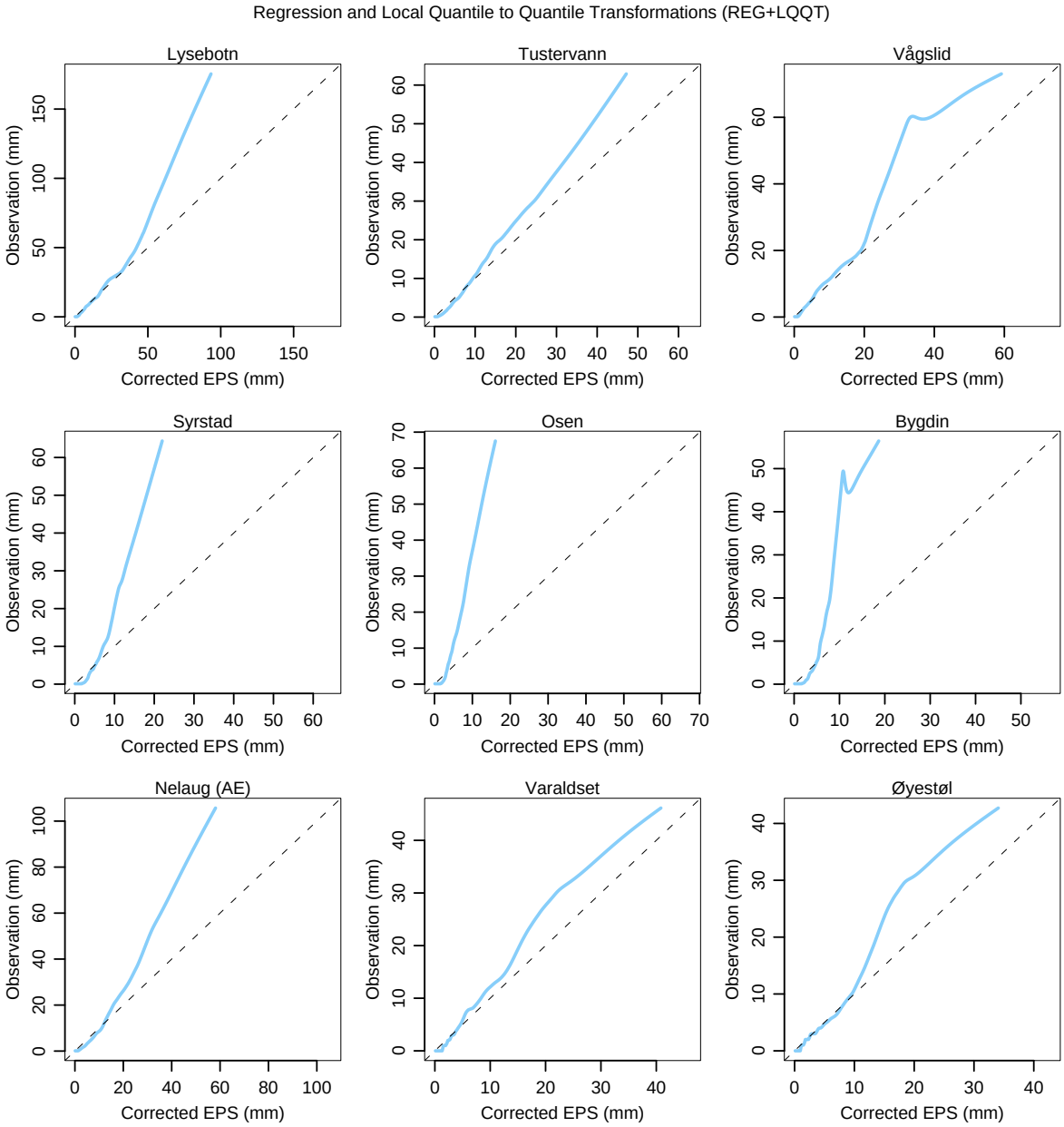

Figure 10: Estimated transformations obtained by least square regression followed by the local quantile-to-quantile transformation method for each of the nine sites.

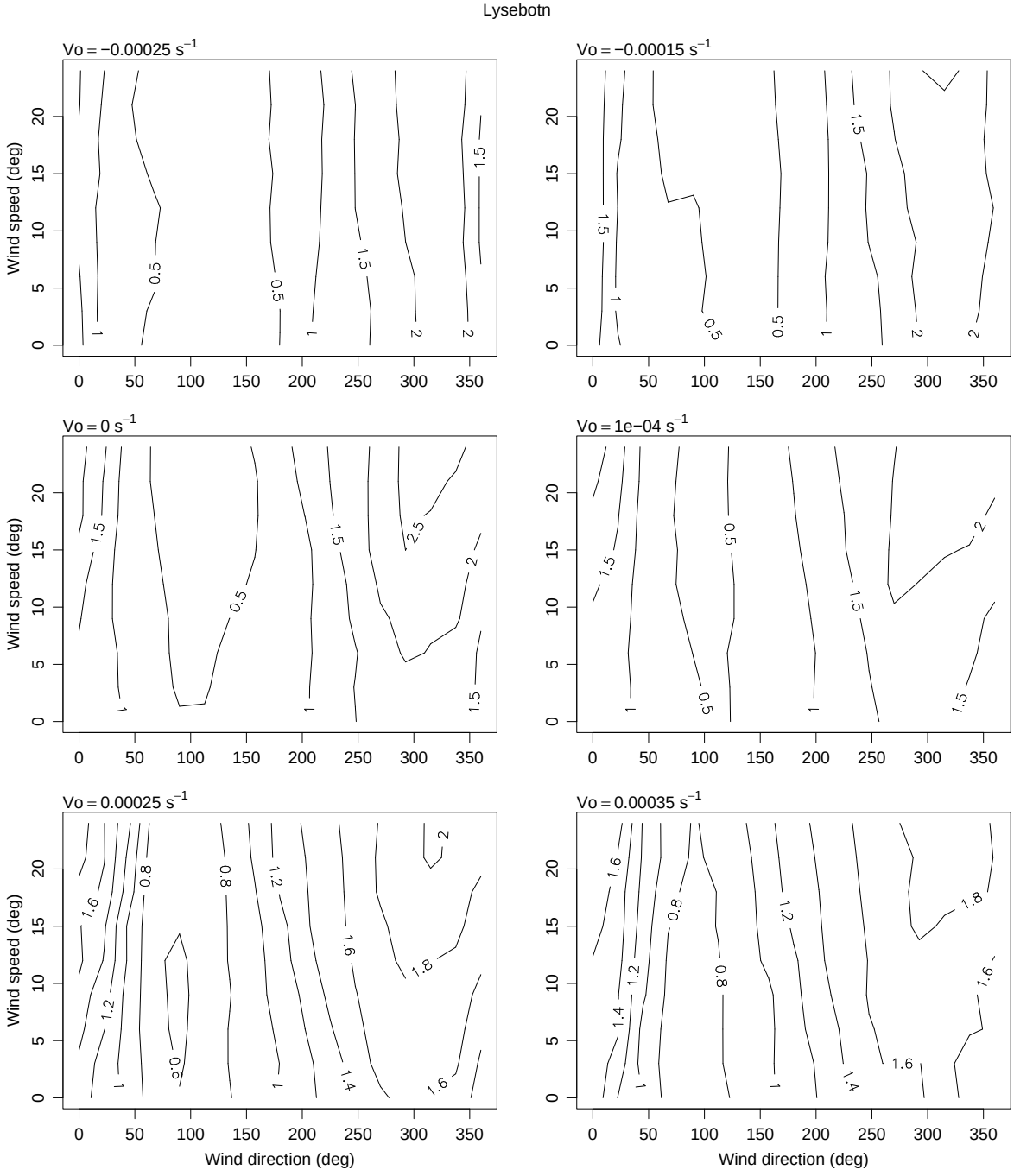

Figure 11: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Lysebotn.

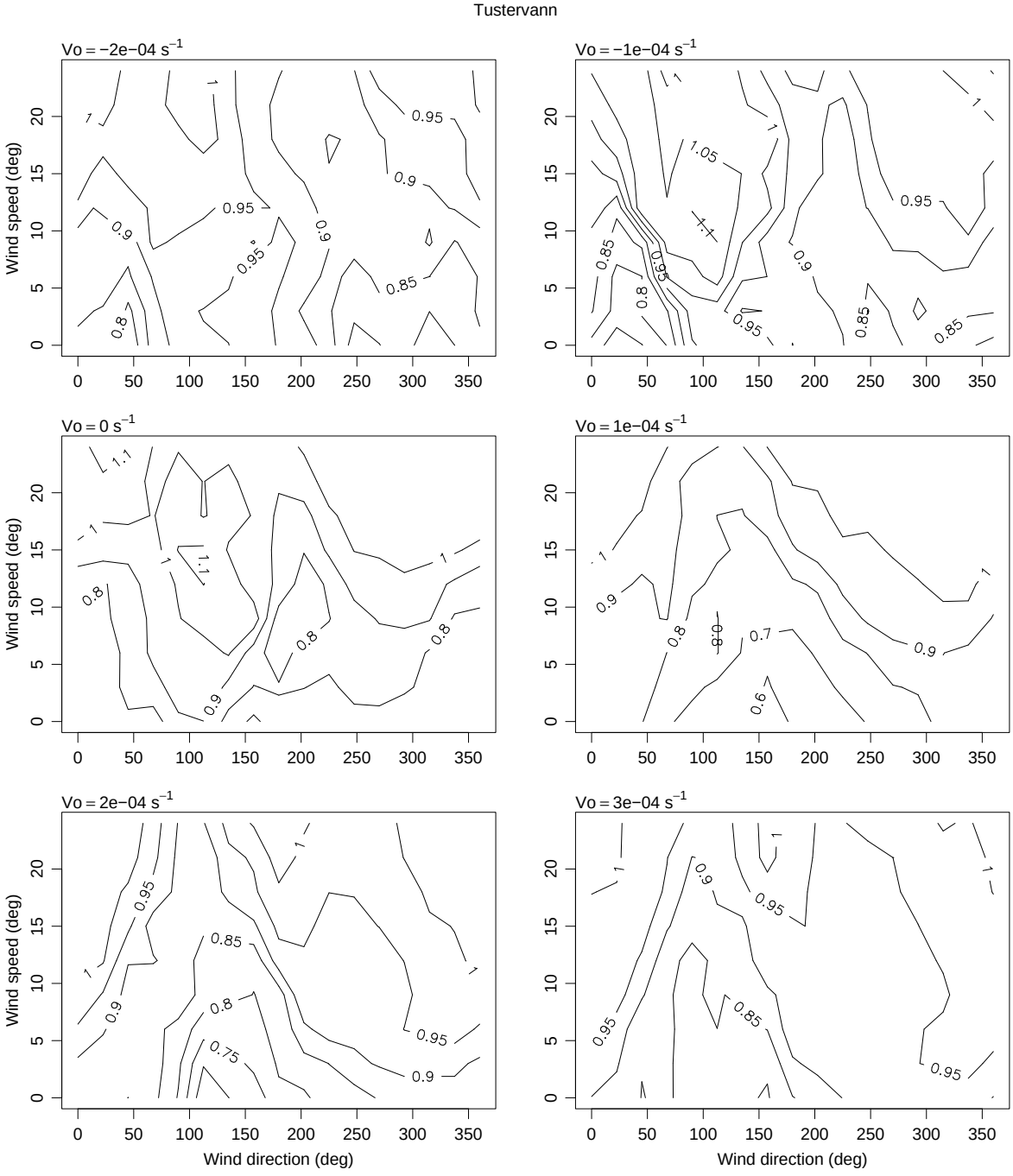

Figure 12: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Tustervann.

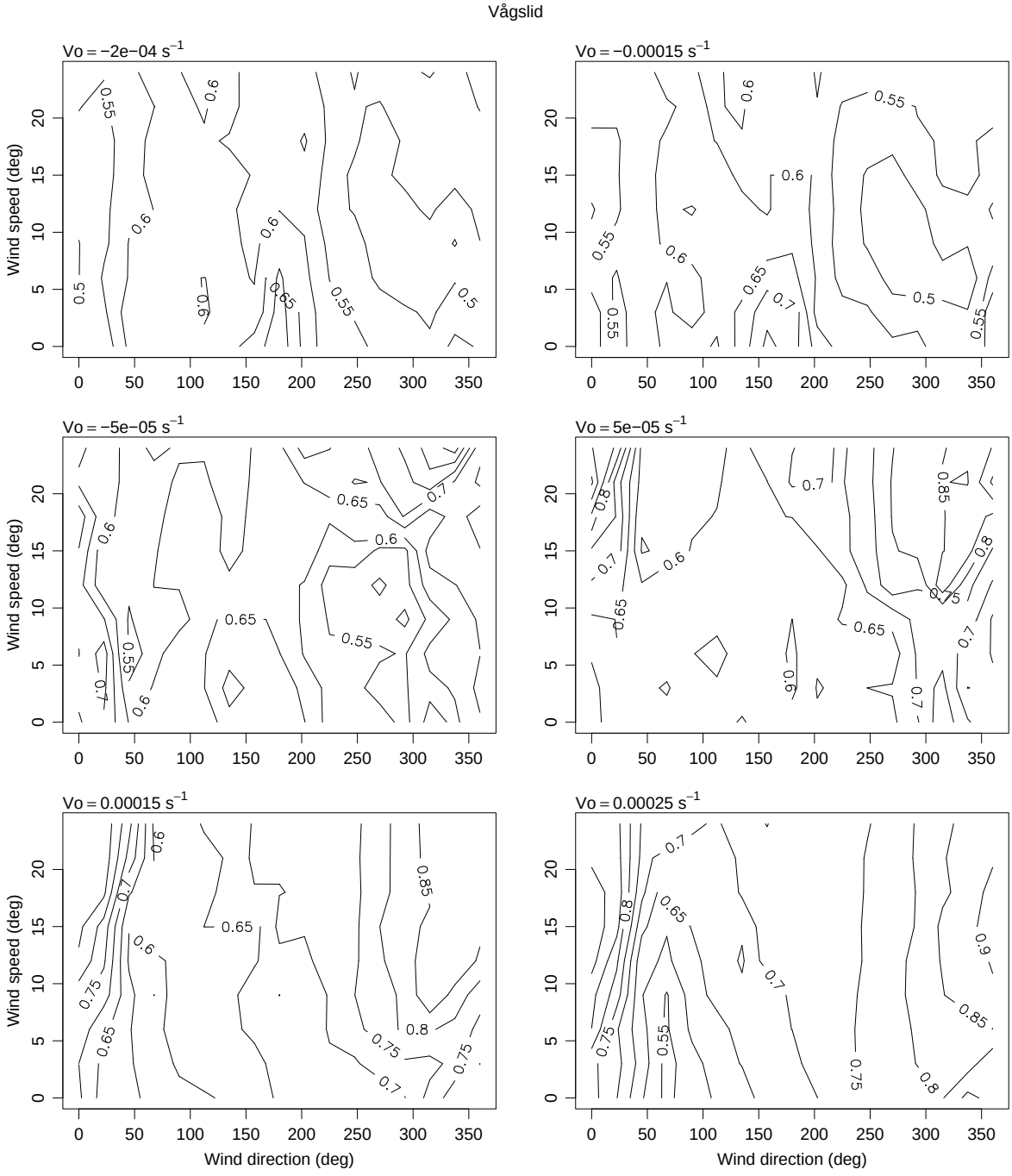

Figure 13: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Vågslid.

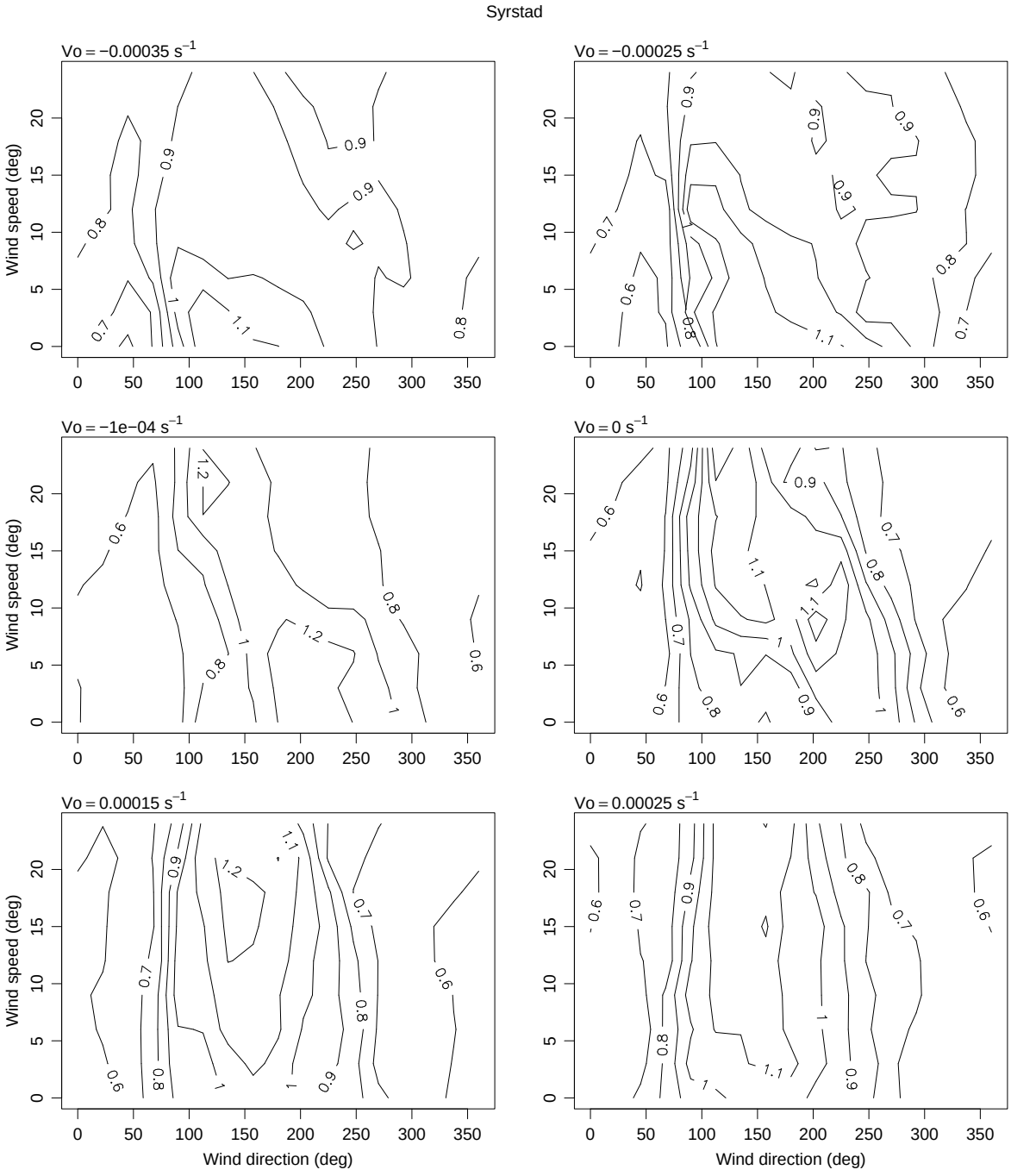

Figure 14: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Syrstad.

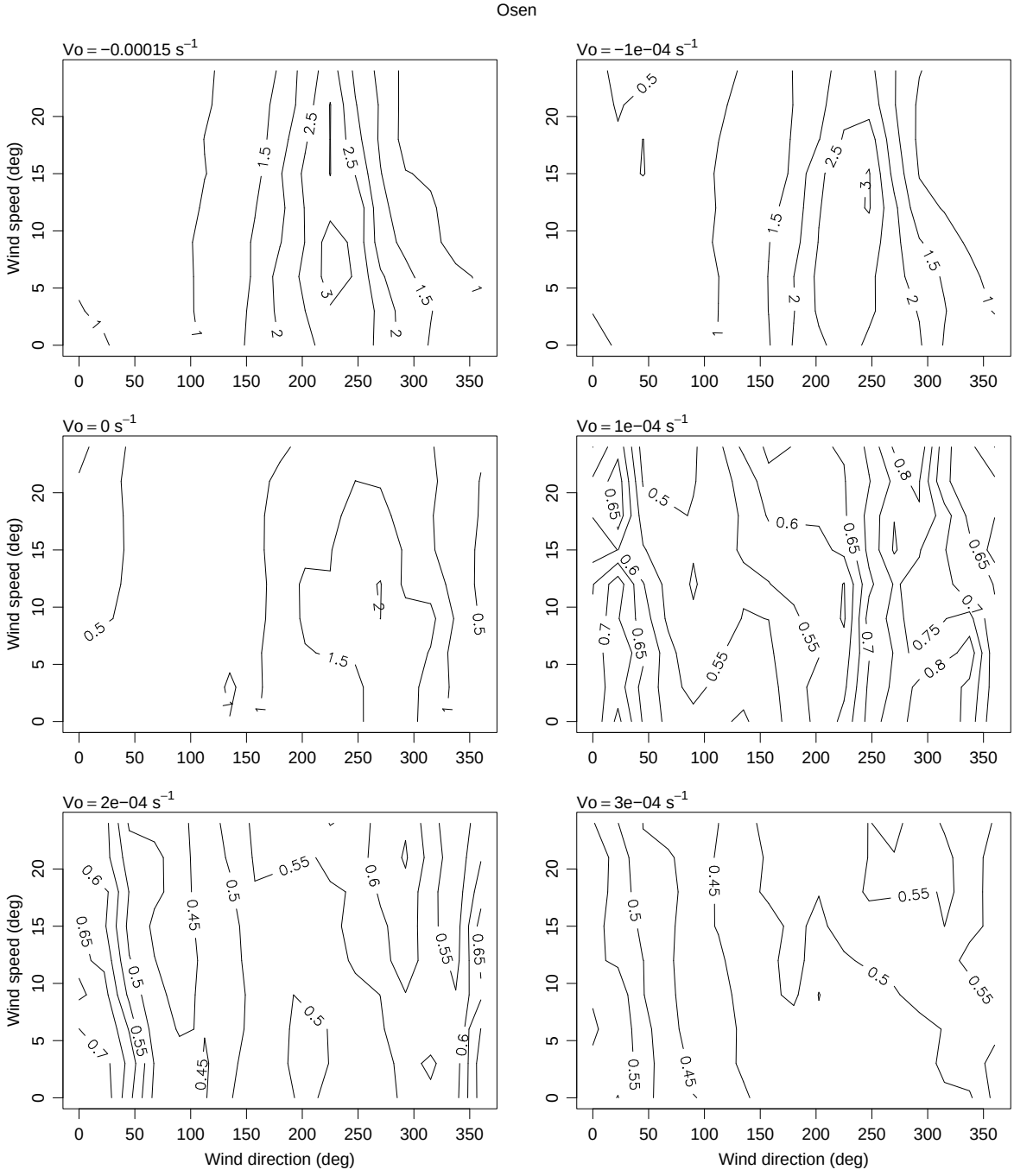

Figure 15: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Osen.

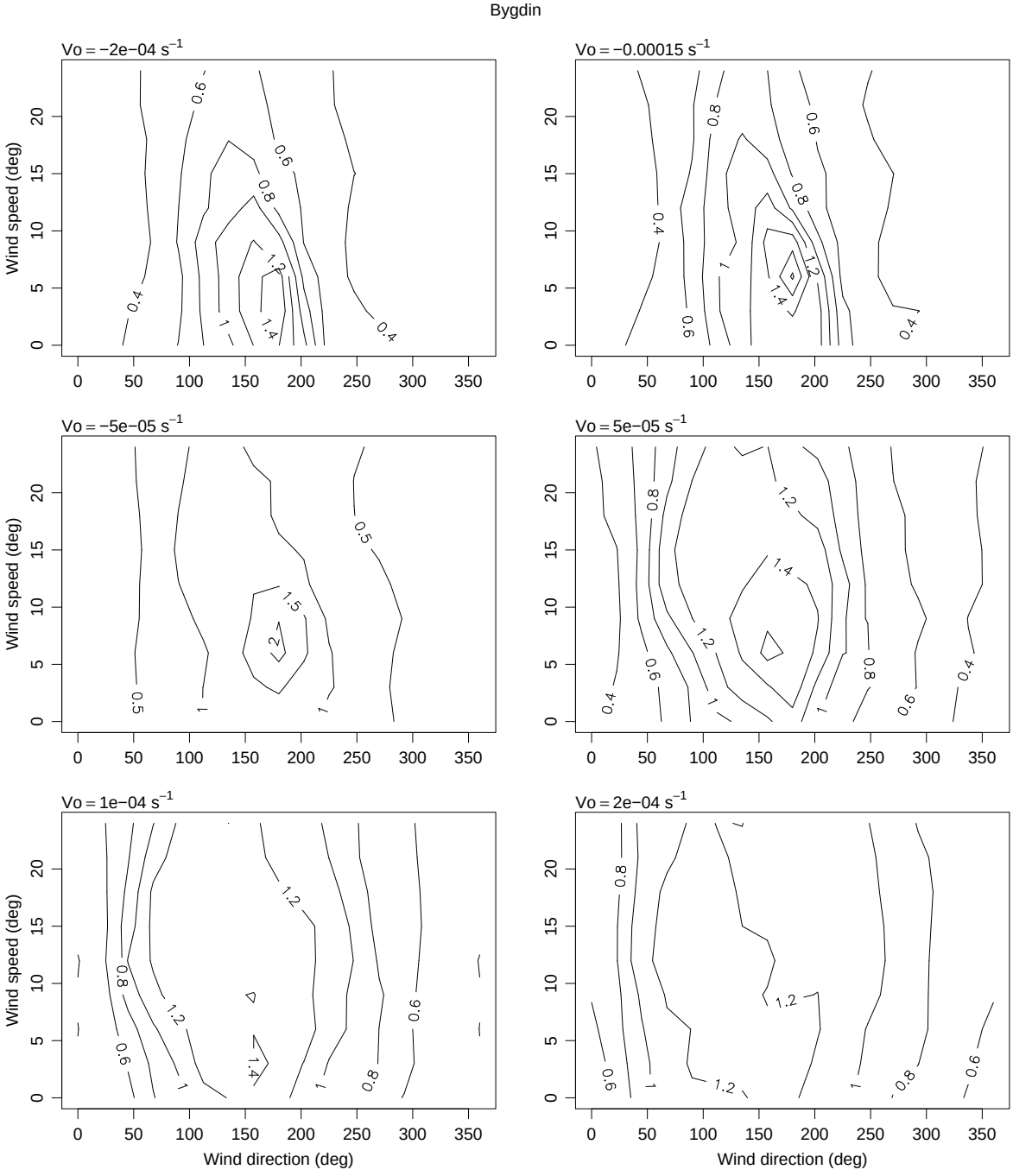

Figure 16: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Bygdin.

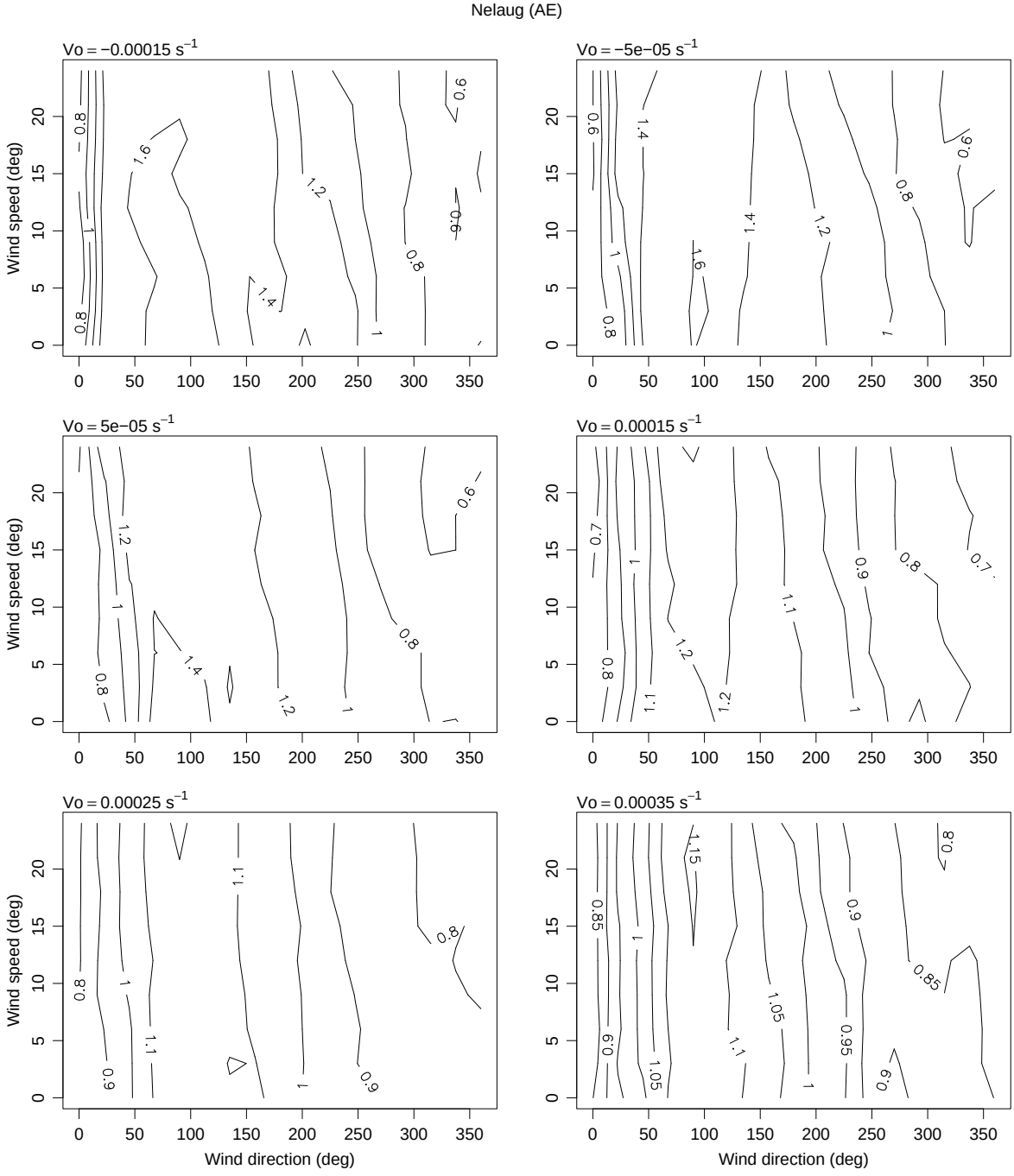

Figure 17: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Nelaug.

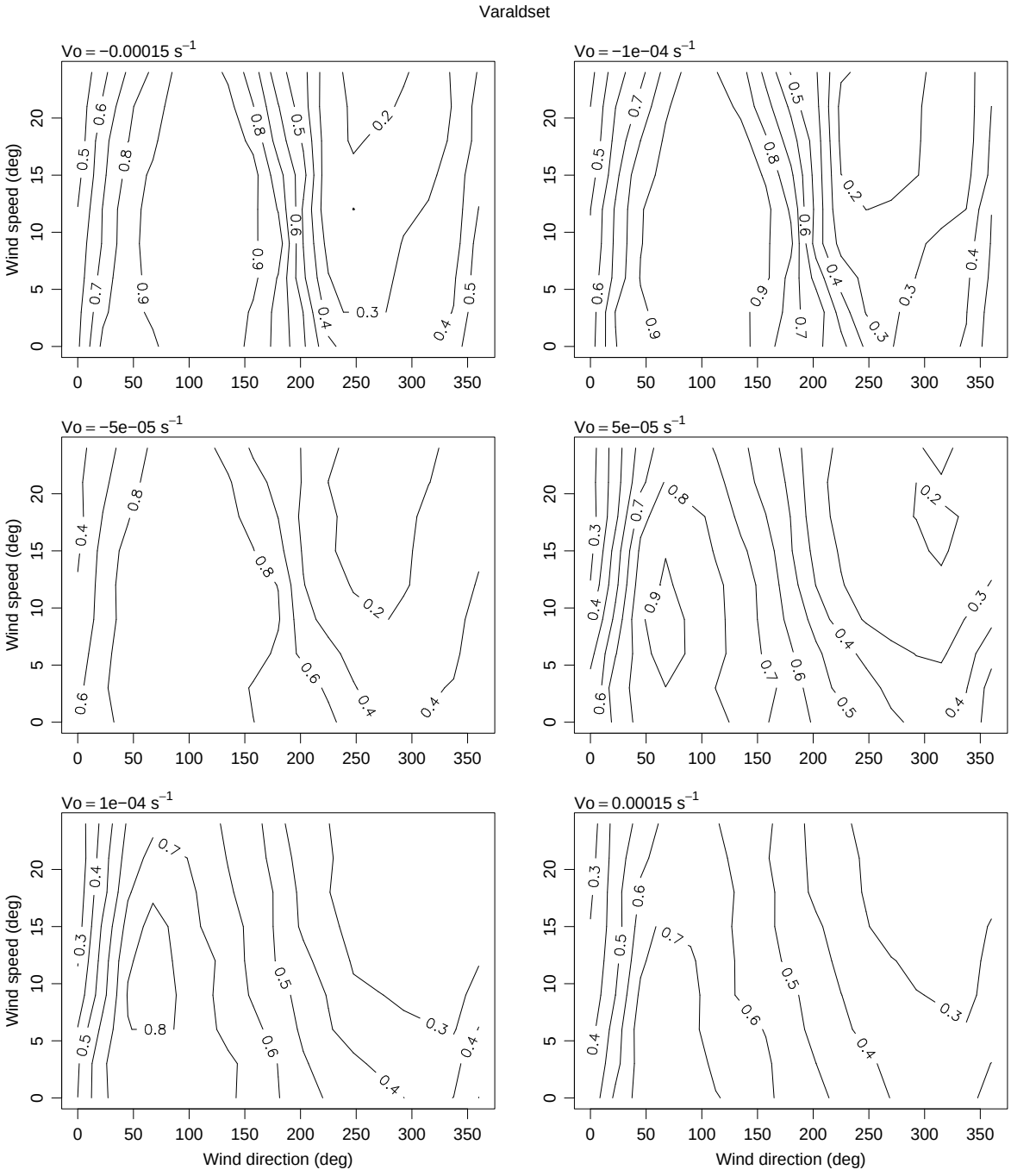

Figure 18: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Varaldset.

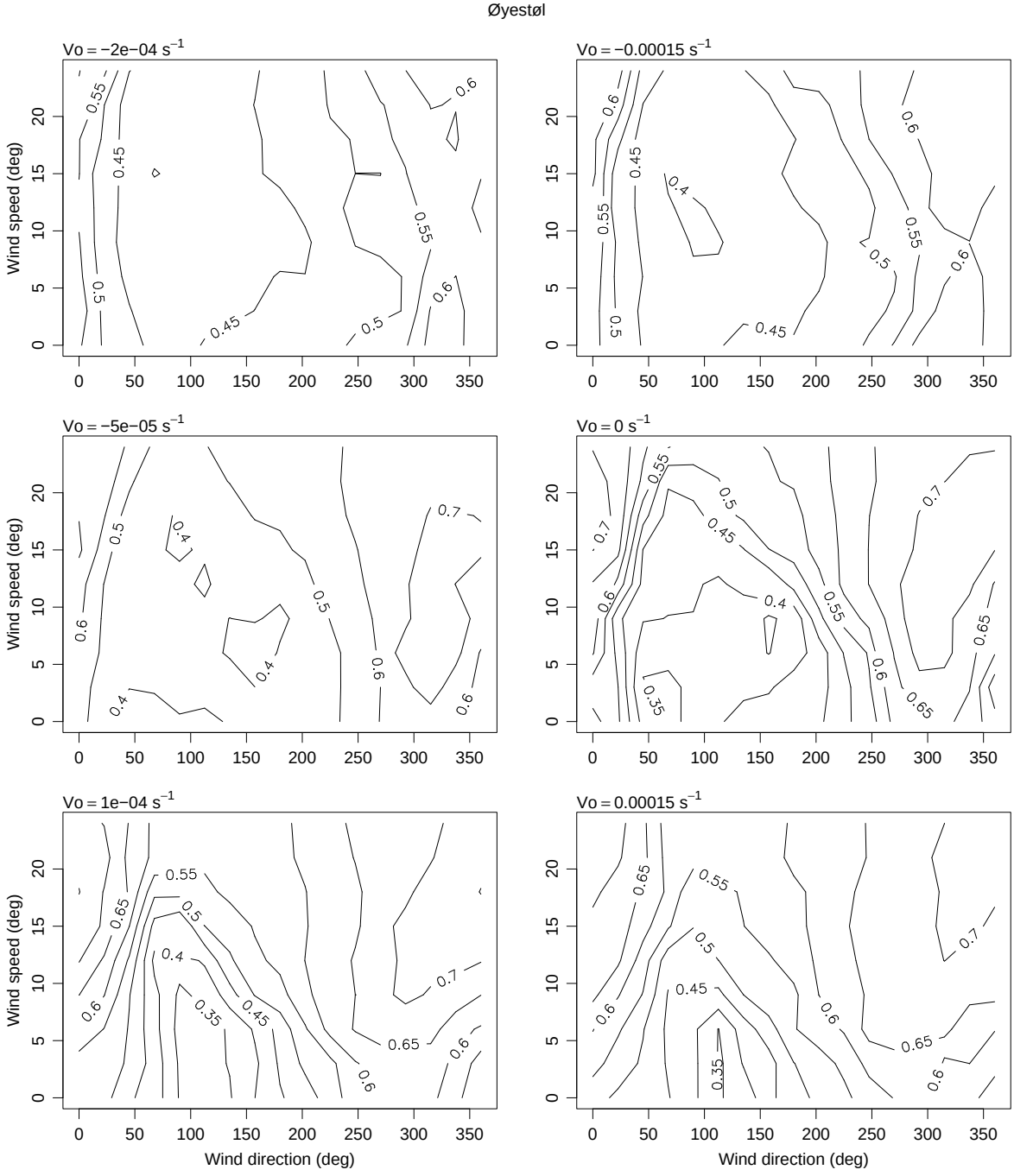

Figure 19: Estimated scaling factors as a function of wind direction, wind speed and relative vorticity (Vo) at Øyestøl.