Probabilistic Forecasting of Wind and Wind Power

PhD Thesis

in Atmospheric Sciences

Institute of Meteorology and Geophysics
Leopold-Franzens-University Innsbruck

by

Jakob W. Messner

Advisors

Georg J. Mayr, Achim Zeileis, and Jochen Bröcker

Innsbruck, October 2013

Suggested reviewing committee:

1. Pierre Pinson, Technical University of Denmark, DTU Elektro
2. Tom Hamill, NOAA Earth System Research Laboratory, Physical Science Division

Abstract

Wind power has experienced rapid growth over the past decades and has become an important part of many power systems. To cope with the volatility of wind and wind power, grid operators and managers strongly demand accurate wind and wind power predictions. Probabilistic forecasts that also allow them to estimate the forecast uncertainty are often most valuable. This thesis aims at testing, developing, and improving statistical methods to post-process numerical weather prediction (NWP) ensembles for probabilistic wind and wind power forecasts. In the first part, a novel approach is proposed to address the conversion problem from wind to wind power. This approach uses an inverse power-curve transformation and censored regression models and showed similar and for small training data sets even better forecast performance than more complex benchmark models. In the second part, ensemble post-processing with extended logistic regression is improved with a new approach to utilize ensemble spread information. Finally, the third part compares extended logistic regression with closely related ordered and censored regression models. Results from a case study with wind and precipitation data suggest that the choice of the optimal regression model strongly depends on the intended application.

Contents

1

Introduction

1.1 General introduction

Satisfying the steadily increasing demand for energy while simultaneously reducing the emissions of carbon dioxide is a major challenge for the 21st century. Governments around the world put a lot of money and effort into changing from traditional to renewable power sources. In this context, wind power is regarded as one of the most promising alternative power sources. As a consequence of governmental incentives and technical advances a high number of (onshore and offshore) wind power plants have been constructed in the past decades. In 2012 wind power has already reached a portion of 11.4 % of the total European Union's installed power capacity (Wilkes and Moccia 2013). However, the volatility of wind and consequently wind power complicates its integration in the power systems. For a stable power system, it is crucial that energy supply and demand are balanced at any time. Thus, for times with low winds, other power sources have to balance the missing wind power. Unfortunately many types of power plants (e.g., fossil-fueled power plants, nuclear power plants) have to plan their production hours or even days in advance. Accurate forecasts of wind and wind power are therefore essential for an optimal integration of wind energy in the power systems. Probabilistic forecasts that provide information about the expected forecast errors can assist power managers and traders to optimize their decisions.

This thesis aims at improving probabilistic wind and wind power forecasts where the main focus is on improved statistical methods to post-process ensemble forecasts from numerical weather prediction (NWP) models. In the following,

this chapter provides an introduction to wind power prediction in general (1.2) , an introduction to probabilistic and ensemble forecasting (1.3), and a description of the research topics and the outline of this thesis (1.4).

1.2 Wind power prediction

As response to the increasing demand for accurate wind power predictions, a rapidly growing wind power forecasting community has formed in the past decade. As a result, many different approaches to predict wind power have been proposed. A comprehensive overview of these different approaches can be found in Giebel et al. (2011).

For lead times longer than several hours, wind power predictions are usually based on numerical weather prediction (NWP) models that compute future weather based on the current state of the atmosphere and a mathematical description of the atmospheric processes. However, because of imperfectly known atmospheric states and unresolved atmospheric processes, NWP forecasts always exhibit errors. Fortunately some of these errors are systematic and can be corrected with statistical post-processing, often referred to as model output statistics (MOS; Glahn and Lowry 1972).

NWP models usually do not directly predict wind *power*. Thus, NWP based forecasts must involve a conversion from wind to wind power. Actually the kinetic power of wind (P) is a function of air density (ρ), the third power of wind speed (v), and the rotor area (A):

$$P = \frac{1}{2} A \rho v^3 \tag{1.1}$$

However, Betz (1920) showed theoretically that wind turbines can only extract approximately 60 % of this power. Generator efficiency and turbine regulations further limit the power yield of real wind turbines and lead to complex relationships between wind speed and turbine power output. Figure 1.1 shows the power-curve, which describes this relationship for an example turbine. Below a certain cut-in wind speed (v_{CI}) the turbine does not rotate and therefore produces no power. For wind speeds above this cut-in wind speed the power production approximately increases with the third power of wind speed (Equation 1.1). However, above a certain nominal wind speed (v_N) the rotation velocity has to be held constant for security reasons so that the power output is not affected by the wind speed anymore. Finally, the turbine has to be shut down completely when wind exceeds a certain cut-off (v_{CO}) wind speed. Power-curves as shown in Figure 1.1 are usually provided by the turbine manufacturer but can also be derived empirically from observation data (e.g., Cabezon et al. 2004).

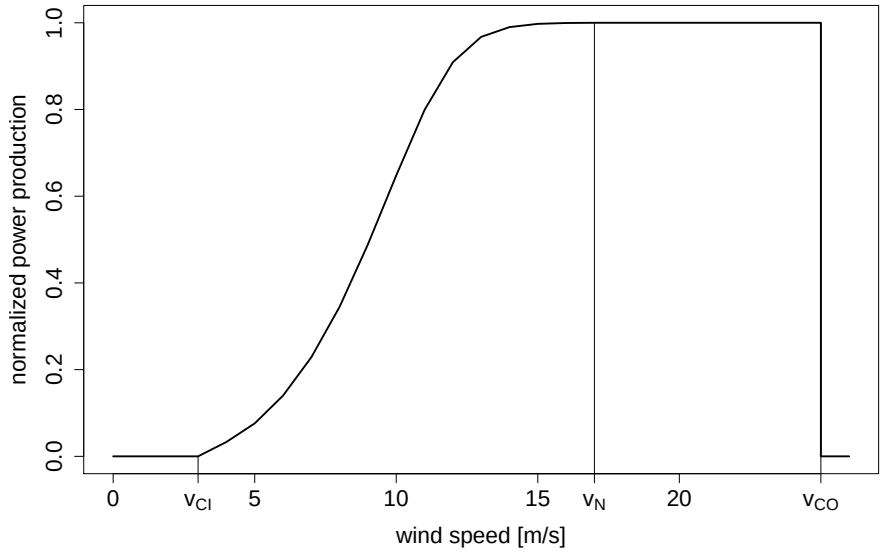

Figure 1.1: Power production (normalized by nominal power) as a function of wind speed. Also plotted are cut-in (v_{CI}), nominal (v_N), and cut-off (v_{CO}) wind speed.

MOS can be used to directly model the power output as a function of NWP wind speed forecasts. Because of the non-linear power-curve (Figure 1.1), usually non-linear and sometimes also non-parametric regression models are employed. Examples of such models are local polynomial regression (Nielsen et al. 2001), local linear quantile regression (Bremnes 2004, 2006), fuzzy neural networks (Pinson 2004), or kernel density estimators (Bremnes 2006; Juban et al. 2007; Bessa et al. 2012a,b).

Non-linear and non-parametric regression models have the disadvantage that many parameters have to be estimated which can lead to unstable parameter estimation, especially if only few data are available for fitting. Furthermore, these models are often hard to interpret and neglect available information about the shape of the power-curve. To overcome these problems, power-curves are often utilized to transform NWP wind *speed* forecasts into wind *power* forecasts. The approximately linear relationship between observed power output and transformed NWP wind speed forecasts then allows to use linear regression models (e.g., Landberg 1999).

However, the limited range (between 0 and nominal power) of both, the observed power production and the transformed NWP forecast complicates parametric distribution assumptions that are usually implied in linear regression models. Thus, nonlinear and nonparametric regression models often still out-

perform simpler parametric linear models (e.g., Nielsen et al. 2006; Møller et al. 2008).

1.3 Forecast uncertainty estimation

Although NWP forecasts are steadily improved (Simmons and Hollingsworth 2002) they are still far from being perfect and always involve some level of uncertainty (Lorenz 1996). Probabilistic forecasts that allow decision makers to estimate this uncertainty can therefore be of high value. Several studies showed that probabilistic forecasts can considerably increase the revenue of wind power traders on day ahead markets. (e.g., Roulston et al. 2001; Bremnes 2004; Zugno et al. 2012).

Most statistical models already provide uncertainty informations which are mostly based on parametric distribution assumptions. In addition, specific non-parametric regression models have been proposed for wind power quantile or interval forecasts. Examples are quantile regression (Bremnes 2004, 2006; Nielsen et al. 2006; Møller et al. 2008), kernel density estimators (Bremnes 2006; Juban et al. 2007; Bessa et al. 2012a,b), or a resampling approach (Pinson and Kariniotakis 2004).

Forecast uncertainty can also be estimated based on physical considerations. Imperfect initial conditions and model formulations are two major error sources of NWP models. Ensemble forecasts try to consider these error sources by computing different NWP model forecasts with slightly perturbed initial conditions and different model formulations. The resulting individual forecasts are then presumably span the range of possible outcomes (Lorenz 1996). Several studies have shown the advantage of ensemble forecasts for wind power predictions (e.g., Roulston et al. 2001; Roulston 2003; Giebel et al. 2005; Nielsen et al. 2007; Pinson and Madsen 2009). Unfortunately the perturbed initial conditions of the different ensemble members usually do not perfectly represent initial condition uncertainty (Hamill et al. 2003; Wang and Bishop 2003) and some structural deficiencies in the NWP models are also not accounted for. Thus, ensemble forecasts do not perfectly represent the full forecast uncertainty of NWP models. Nevertheless, ensemble forecasts can contain valuable information about forecast uncertainty and can be statistically post-processed to achieve well-calibrated probabilistic forecasts.

In the past decades a variety of different statistical ensemble post-processing methods have been proposed. For example Roulston and Smith (2003), Wang and Bishop (2005), and Pinson and Madsen (2009) proposed to dress individual ensemble members with historical forecast error distributions. Bayesian model av-

eraging (Raftery et al. 2005) is a closely related method with differently weighted ensemble members. Several papers also suggested to use linear regression models with heteroscedastic error distributions (ensemble-MOS Gneiting et al. 2005; Thorarinsdottir and Gneiting 2010; Scheuerer 2013). Ensemble copula coupling (e.g., Schefzik et al. 2013; Schuhen et al. 2012; Pinson 2012) is a recent method to achieve calibrated and coherent forecast ensembles and logistic regression is another important ensemble post-processing method that is well-suited for binary predictands.

Comparisons of some of these and other methods (Wilks 2006a; Wilks and Hamill 2007) showed the good performance of logistic regression. Recently, Wilks (2009) extended logistic regression by including the predictand thresholds as additional predictor variable, which allows derivation of full continuous predictive distribution. Since then, this extended logistic regression has been used frequently, mainly to post-process precipitation ensemble forecasts (Schmeits and Kok 2010; Ruiz and Saulo 2012; Roulin and Vannitsem 2012; Hamill 2012; Ben Bouallègue 2013; Scheuerer 2013).

1.4 Research topics and outline

The main goal of this thesis is to develop, test, and improve statistical ensemble post-processing methods for probabilistic wind and wind power predictions. To reach this overarching goal this thesis addresses three main issues:

1. The first part of this thesis addresses the conversion problem from wind to wind power. In the proposed approach the power outputs are transformed into wind speeds by using the inverse of the power curve function. With this transformation, information about the limited range of power production can easily be exploited with censored models, so that simple linear regression models with parametric distribution assumptions can be used.

2. Although extended logistic regression has mainly been used for ensemble post-processing, ensemble spread information was often neglected because it did not improve the forecasts. The second part of this thesis shows that when included as ordinary predictor variable in extended logistic regression the ensemble spread only affects the location (mean) but not the dispersion (variance) of the predictive distribution. However, the ensemble spread is generally expected to mainly contain information about the forecast uncertainty which in turn should be directly related to the dispersion of the predictive distribution. Based on this finding a heteroscedastic extended logistic regression approach is proposed where the ensemble spread can be

directly used to predict the dispersion of the predictive distribution.

3. Ordered and censored logistic regression are popular regression models from statistics and econometrics that are very similar to extended logistic regression but have not received much attention in meteorology so far. The third part of this thesis compares standard, ordered, extended, and censored logistic regression and tests their suitability for different applications.

Each of these issues is treated in a separate scientific article. The remainder of this thesis consists of these three articles and an overall summary and conclusion (5).

2

Paper I

Messner, J. W., A. Zeileis, J. Broecker, and G. J. Mayr, 2013: Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. *Wind Energy*, in press.

Probabilistic wind power forecasts with an inverse power curve transformation and censored regression[1]

Jakob W. Messner[2] and Georg J. Mayr
Institute of Meteorology and Geophysics, University of Innsbruck, Austria

Jochen Bröcker
Department of Mathematics and Statistics, University of Reading, United Kingdom

Achim Zeileis
Department of Statistics, University of Innsbruck, Austria

ABSTRACT

Forecasting wind power is an important part of a successful integration of wind power into the power grid. Forecasts with lead times longer than 6 hours are generally made by using statistical methods to postprocess forecasts from numerical weather prediction systems. Two major problems that complicate this approach are the nonlinear relationship between wind speed and power production and the limited range of power production between zero and nominal power of the turbine. In practice, these problems are often tackled by using nonlinear nonparametric regression models. However, such an approach ignores valuable and readily available information: the power curve of the turbine's manufacturer. Much of the nonlinearity can be directly accounted for by transforming the observed power production into wind speed via the inverse power curve so that simpler linear regression models can be used. Furthermore, the fact that the transformed power production has a limited range can be taken care of by employing censored regression models.

In this study, we evaluate quantile forecasts from a range of methods: (a) using parametric and nonparametric models, (b) with and without the proposed inverse power curve transformation, and (c) with and without censoring. The results show that with our inverse (power-to-wind) transformation, simpler linear regression models with censoring perform equally or better than nonlinear models with or without the frequently used wind-to-power transformation.

[1] in press in *Wind Energy*

[2] *Corresponding author address:* Institute of Meteorology and Geophysics, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria. E-mail: jakob.messner@uibk.ac.at

2.1 Introduction

The importance of wind energy has increased significantly in the past decades. In 2011 approximately 21% of installed power capacity in Europe was from wind power (Wilkes et al. 2012). One problem of integrating wind power into the electricity grid is the volatility of wind speed and consequently of power production. Prediction of power production is therefore crucial for energy trading and management. In this context, probabilistic forecast methods have been receiving increased attention recently because of their higher value in decision making when compared to single value (point) forecasts (Pinson et al. 2007; Roulston and Smith 2003; Bremnes 2004). Probabilistic forecasts can for example be quantile or interval forecasts, full predictive distributions, or risk indices in addition to point forecasts.

The general approach to make probabilistic power production forecasts with lead times ≥ 6 hours is to statistically postprocess forecasts (mainly wind speed forecasts) from numerical weather prediction (NWP) models (Giebel et al. 2011). In the atmospheric sciences, this approach is termed model output statistics (MOS; Glahn and Lowry 1972). However, standard linear regression analysis, as typically used for MOS, is complicated by two major problems:

1. The relationship between wind speed and power production is clearly nonlinear (see Figures 2.1 and 2.2) .
2. The range of power production is limited between zero and nominal power so that typical parametric distribution assumptions (e.g., Gaussian) are inappropriate.

To overcome these problems, nonlinear and often also nonparametric regression methods are used frequently in the literature. For example, a variety of nonlinear quantile regression methods have been proposed. Examples are locally weighted quantile regression (Bremnes 2004, 2006), quantile regression with spline basis functions (Nielsen et al. 2006), or a time-adaptive quantile regression (Møller et al. 2008). Other widely used approaches are kernel density estimators and variations of it (Bremnes 2006; Juban et al. 2007; Bessa et al. 2012b,a), ensemble postprocessing, e.g., with kernel dressing (Taylor et al. 2009; Pinson and Madsen 2009) or quantile correction (Nielsen et al. 2004; Giebel et al. 2005), or adaptive resampling (Pinson and Kariniotakis 2010). The disadvantages of such nonparametric nonlinear models are that generally a large number of parameters have to be estimated and therefore these estimations can be unstable, especially in cases where few data are available. Furthermore, the resulting models are sometimes hard to interpret and, more importantly, neglect the available information about the form of the power curve and the censoring.

Therefore we propose a new (line of) approach(es):

1. Transform the observed power observations into wind speed observations prior to MOS regression modeling by using the inverse of the power curve function. Note that this transforms the limited range from zero to nominal power into the limited range from cut-in wind speed to nominal wind speed.

2. Exploit the information about this limited range by using censored models in "wind space" where typically much simpler (more) linear regressions can be used and parametric distributions work well.

Figure 2.4 shows the relationship between power observations, transformed with the inverse power curve on the y-axis and NWP wind speed forecasts on the x-axis. Clearly, this seems to be almost linear and just the censoring of the transformed power observations at cut-in and nominal wind speed has to be accounted for in a regression model. While such censored regression techniques are not very frequently used for MOS, they are among the standard regression models in statistics and econometrics and easily available in many software statistics packages. Thus we can obtain probabilistic forecasts in "wind space" with a relatively simple model and then employ the power curve again to transform these to probabilistic power production forecasts.

We are not the first to suggest usage of the known power curve to address the nonlinearity issue. However, previous approaches employed the power curve itself rather than its inverse to transform the NWP wind speed forecasts into power forecasts prior to regression modeling (Lange 2005; Nielsen et al. 2004; Roulston 2003). While this is also very easy to carry out (see Figure 2.3 for an example), it has a crucial disadvantage: In the steep parts of the power curve, errors in the NWP wind speed forecasts are strongly amplified while errors of low and high NWP wind speed forecasts are suppressed. Hence, the resulting relationship between the (wind-to-power) transformed NWP wind speed forecasts and observed power production exhibits strong heteroskedasticity which leads to less reliable estimates in regression models. Note the higher variance in the center of Figure 2.3 as compared with the lower variance on the left and right side. In contrast, the inverse power-to-wind transformed relationship in Figure 2.4 has a rather low and stable variance (only limited by censoring at cut-in and nominal speed).

In this study, we demonstrate how both parametric and nonparametric censored (linear) regression models can be employed for inverse power curve transformed data (i.e., in wind space). The resulting models are assessed and compared with previously suggested approaches for untransformed data as well as

wind-to-power transformed data (i.e., in power space), showing that in many situations we can get similar or even better performance from models that are easier to compute and interpret. As observation data, we use 3 years of wind turbine data from a turbine located in Austria. As NWP forecasts, high resolution and ensemble forecasts of wind in different heights from the European Centre for Medium-Range Weather Forecasts (ECMWF) are employed.

The remainder of the paper is organized as follows: In Section 2.2, the data used for testing the transformations and models are described briefly. The regression models are introduced in Section 2.3. The verification measures are specified in Section 2.4 and the corresponding results are shown in Section 2.5. Finally, a conclusion of the paper is provided in Section 2.6.

2.2 Data

As observation data, we utilize power production data from a wind turbine in eastern Austria with a nominal power of 2000kW. Measurements with 10 minute temporal resolution are available from 2006 to 2009. Data values when the turbine was off because of maintenance are removed.

As input for the statistical models, we use NWP forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). In particular, we use wind speed forecasts, linearly interpolated from neighbouring model levels to turbine hub height, as this has been shown to be the best predictor from ECMWF for wind speed on wind turbines (Drechsel et al. 2012). No further variables (such as wind direction or air density) are added because they do not improve forecasts significantly for the data considered. To capture heteroskedasticity (i.e., inhomogeneous, input dependent standard deviation of the observations) some of our models additionally employ the 10 meter wind speed ensemble standard deviation form the ECMWF ensemble prediction system (EPS). To combine the observation data (with temporal resolution of 10 minutes) with the NWP data (with resolution of 3 hours), means of the observation data are computed for 1 hour around the times for which forecasts are available.

Thus for each lead time, 1340 forecast-observation pairs are available. For lead time of 24 hours, the data is plotted in Figure 2.2. Note that all used ECMWF forecasts are initialized at 00UTC. 12 and 36 hour forecasts are therfore always for midday while 24 and 48 hour forecasts are always for midnight.

In the next sections the following notations are used:

n**:** Number of forecast-observation pairs.

p_i**:** Power production; $i = 1, ..., n$.

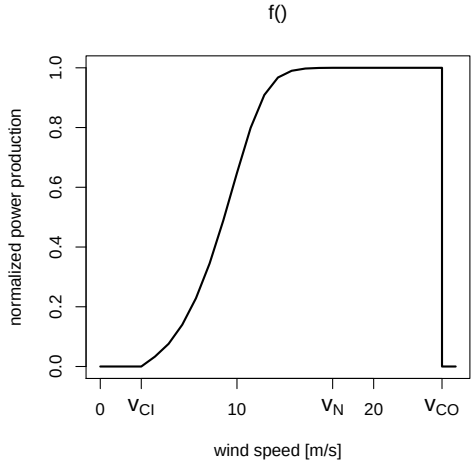

Figure 2.1: Power curve function $f()$ of the turbine manufacturer: Power production by wind speed.

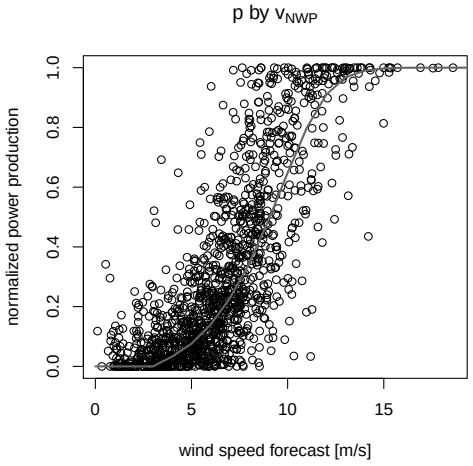

Figure 2.2: Normalized power production (black points) by ECMWF wind speed forecasts with power curve (gray line).

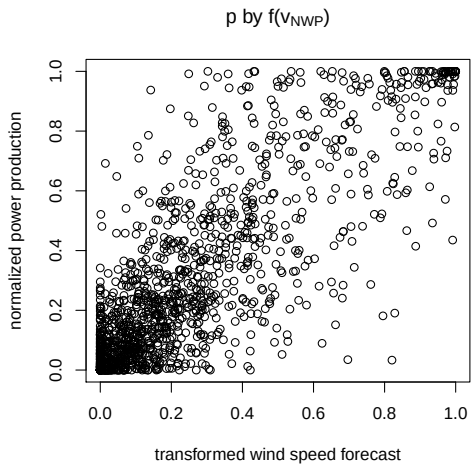

Figure 2.3: Power curve transformation (wind to power): Observed power production by transformed ECMWF wind speed forecasts.

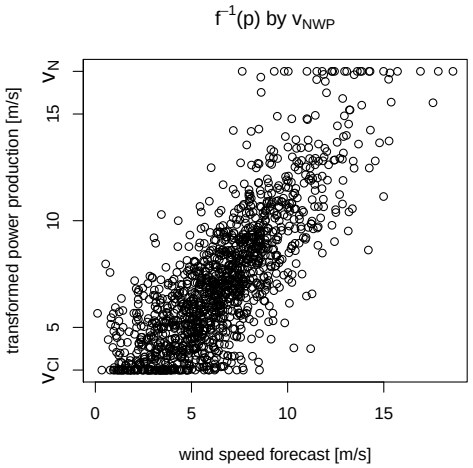

Figure 2.4: Inverse power curve transformation (power to wind): Transformed power production by ECMWF wind speed forecasts.

v_i^***:** Wind speed; $i = 1, ..., n$.

v_{CI}**:** Cut-in wind speed (wind speed where turbine starts to rotate).

v_N**:** Nominal wind speed (wind speed where turbine reaches maximum power).

$f()$**:** Power curve function given by the turbine manufacturer; $f(v_i^*) = p_i$ (see Figure 2.1).

$v_i = f^{-1}(p_i)$**:** Inverse-transformed power production (see also Equation 2.1).

$\mathbf{x}_i, \mathbf{z}_i$**:** Vectors of input variables (NWP forecasts); $i = 1, ..., n$.

$q_\pi(y_i|\mathbf{x}_i)$**:** π-quantile of y_i given the regressor variables $\mathbf{x}_i$.

Note that the inverse-transformed power production (v_i) can be interpreted as wind speed censored at cut-in and nominal wind speed (see Figure 2.4). That means:

$$v_i = f^{-1}(p_i) = \begin{cases} v_{CI} & v_i^* \leq v_{CI} \\ v_i^* & v_{CI} < v_i^* < v_N \\ v_N & v_i^* \geq v_N \end{cases} \tag{2.1}$$

Note that an inverse-transformed power production of $v_i = v_{CI}$ can also occur at very high wind speed when the turbine has to be switched off in order to avoid damages. However, switching off the turbine because of too high wind speed did never happen in our data and is therefore not considered in the following.

2.3 Regression models

To obtain probabilistic forecasts of power production, we consider a range of different regression models that lead either to conditional quantiles or full predictive distributions (from which conditional quantiles can easily be extracted). More formally, all models yield predictions of specific quantiles $q_\pi(p_i|\mathbf{x}_i)$ of power production p_i given a vector of regressor variables $\mathbf{x}_i$ (e.g., forecasts of wind speed etc.). We divide the models into parametric and nonparametric models. All models except some benchmark models are estimated in wind space. That means that quantiles $q_\pi(v_i^*|\mathbf{x}_i)$ of wind speed given some regressor variables are first estimated. Subsequently they are transformed to quantiles of transformed power by considering cut-in and nominal wind speed of the turbine and finally transformed into quantiles of power production by employing the power curve of the turbine:

$$q_\pi(v_i|\mathbf{x}_i) = \min(v_N, \max(v_{CI}, q_\pi(v_i^*|\mathbf{x}_i))) \tag{2.2}$$

$$q_\pi(p_i|\mathbf{x}_i) = f(q_\pi(f^{-1}(p_i)|\mathbf{x}_i)) = f(q_\pi(v_i|\mathbf{x}_i)) \tag{2.3}$$

2.3.1 Parametric models

For parametric models, it is assumed that the response follows a specific distribution and here the normal (or Gaussian) distribution is used. If such an assumtion is appropriate these models are easy to estimate and with every forecast a full

predictive ditribution is given. Arbitrary quantiles are very easy to compute by inverting this distribution. The main disadvantage of parametric models is that it is sometimes difficult to find an appropriate parametric distribution.

Tobit model

The tobit model was first introduced by Tobin (1958) and is a widely used linear model for censored data. For this model, it is assumed that the true wind speed v_i^* follows a normal distribution with a mean μ_i that depends linearly on some input variables $\mathbf{x}_i$ and typically a constant variance $\sigma_i = \gamma$:

$$v_i^* \sim N(\mu_i, \sigma_i^2) \tag{2.4}$$

$$\mu_i = \mathbf{x}_i^\top \beta \tag{2.5}$$

$$\sigma_i = \gamma \tag{2.6}$$

However, as outlined above, the wind speed obtained by transforming the observed power production (v_i) is censored at cut-in and nominal wind speed (Equation 2.1). Thus the coefficients β and σ are not estimated with standard least squares regression but with maximum likelihood estimation with the likelihood function

$$L(\beta, \gamma | v_i, \mathbf{x}_i) =$$

$$\prod_{i=1}^{n} f(v_i|\mathbf{x}_i, \beta, \gamma)^{I(v_{CI}<v_i<v_N)} P(v_i = v_{CI}|\mathbf{x}_i, \beta, \gamma)^{I(v_i=v_{CI})} P(v_i = v_N|\mathbf{x}_i, \beta, \gamma)^{I(v_i=v_N)} \tag{2.7}$$

where the indicator function $I(a)$ is 1 if the argument a is true and is 0 if it is not. Furthermore

$$P(v_i = v_{CI}|\mathbf{x}_i, \beta, \gamma) = P(v_i^* \leq v_{CI}|\mathbf{x}_i) = \Phi\left(\frac{v_{CI} - \mathbf{x}_i^\top \beta}{\sigma_i}\right) \tag{2.8}$$

$$P(v_i = v_N|\mathbf{x}_i, \beta, \gamma) = P(v_i^* \geq v_N|\mathbf{x}_i) = 1 - \Phi\left(\frac{v_N - \mathbf{x}_i^\top \beta}{\sigma_i}\right) \tag{2.9}$$

$$f(v_i|\mathbf{x}_i, \beta, \gamma) = \frac{1}{\sigma_i}\phi\left(\frac{v_i - \mathbf{x}_i^\top \beta}{\sigma_i}\right) \tag{2.10}$$

where Φ and ϕ are the cumulative distribution function and the probability density function of the standard normal distribution, respectively. With this model, conditional quantile forecasts for v_i^* can be computed with

$$q_\pi(v_i^*|\mathbf{x}_i) = \mathbf{x}_i^\top \beta + \sigma_i \Phi^{-1}(\pi). \tag{2.11}$$

Heteroskedastic tobit model

The standard tobit model assumes a constant residual variance σ_i over all $i = 1, \ldots, n$. This assumption can be relaxed with an additional regression equation for the standard deviation σ_i. Thus, Equation 2.6 is generalized to

$$\log(\sigma_i) = \mathbf{z}_i^\top \gamma \tag{2.12}$$

where $\mathbf{z}_i$ is an additional vector of input variables, not necessarily equal to $\mathbf{x}_i$. The log link is used to assure positive variances. All remaining Equations 2.4–2.11 can still be applied as before.

The heteroskedastic version of the tobit model is used less frequently in the literature. However, for example, Thorarinsdottir and Gneiting (2010) proposed a closely related model with the main difference being that the parameters are estimated by minimizing the continuous ranked probability score (CRPS; Wilks 2006b) instead of maximizing the likelihood function. Their method is a modified version of Gneiting et al. (2005) considering the truncation of wind speed at zero. The method of Gneiting et al. (2005) has proven to perform very well for temperature and precipitation forecasts (Wilks and Hamill 2007).

2.3.2 Nonparametric models

Nonparametric models are more flexible than parametric ones since no distribution of the response has to be assumed. Therefore, they are preferable when no good approximation of the response distribution is known. The price for this flexibility is that only specific quantiles can be estimated and that the model has to be fitted separately for each quantile. If more than one quantile is required, this means that more parameters have to be estimated.

Quantile regression

Similar to the mean in least squares regression, specific quantiles can be estimated with quantile regression. Instead of the quadratic loss function in least squares regression, Koenker and Bassett Jr (1978) proposed to weight residuals above or below the quantile differently, namely

$$\rho_\pi(u) \;=\; \begin{cases} u\pi & \text{if } u \geq 0 \\ u(\pi - 1) & \text{otherwise} \end{cases} \tag{2.13}$$

The π-quantile can be estimated by

$$q_\pi(v_i^* | \mathbf{x}_i) \;=\; \mathbf{x}_i^\top \beta_\pi \tag{2.14}$$

with parameters β_π minimizing

$$\sum_{i=1}^{n} \rho_\pi(v_i - q_\pi(v_i^*|\mathbf{x}_i)) \tag{2.15}$$

Although the censoring of the transformed power production v_i is not considered explicitly in this model, we employ it for comparison to assess the importance of censoring in the regression. Additionally, we use several benchmark models based on quantile regression for observed power production p_i directly as these are used frequently in the wind energy literature (Bremnes 2004, 2006; Nielsen et al. 2006; Møller et al. 2008). See Section 2.3.3 for details on the different models.

Censored quantile regression

As for the parametric models it is also possible to consider censoring with quantile regression. As suggested by Powell (1986), Equations 2.13 and 2.14 still apply and in Equation 2.15, $q_\pi(v_i^*|\mathbf{x}_i)$ is replaced by $q_\pi(v_i|\mathbf{x}_i)$ from Equation 2.2. Note that further approaches to estimate censored quantile regression exist (Portnoy 2003; Peng and Huang 2008; Lin et al. 2012) besides the approach of Powell (1986).

2.3.3 Choice of regressors

In wind space (see Figure 2.4), a simple linear model that uses NWP wind speed forecasts ($v_{NWP,i}$) as the sole regressor is certainly justifiable. However, despite the inverse transformed response, some slight remaining nonlinearities at the lower and upper end appear to remain. These are much weaker than the nonlinearities in the untransformed power-by-wind space (see Figure 2.2) and can be captured very well by a low-dimensional polynomial. Therefore, we consider a number of models that employ not only the linear term $v_{NWP,i}$ but additionally the corresponding squared and cubic terms, i.e., a polynomial of order 3. In addition to these regressors for the mean/quantiles of the predicted wind distribution, the heteroskedastic model also allows for regressors for the standard deviation of the wind distribution. A natural candidate is the ensemble standard deviation of the 10 meter wind speed ($\sigma(\mathbf{v}_{EPS,i})$).

Combining these ideas, we consider a number of models listed in Table 2.1. The tobit model with NWP wind speed forecasts as single regressor variable is the simplest model and already produces a reasonable fit of the data (see *tobit1* in Figure 2.5). Adding the 2nd and 3rd powers to the regressor improves the fit somewhat (*tobit3*). Neither polynomials with higher powers nor the inclusion of further NWP variables (e.g., air density, 10m wind speed ensemble mean, sine and cosine of wind direction) as regressors lead to further significant improvements

Model		Response	Regressors
tobit1	Tobit model	v_i^*	$\mathbf{x}_i = v_{NWP,i}$
tobit3	Tobit model	v_i^*	$\mathbf{x}_i = (v_{NWP,i}, v_{NWP,i}^2, v_{NWP,i}^3)$
htobit1	Heteroskedastic tobit model	v_i^*	$\mathbf{x}_i = v_{NWP,i}$, $\mathbf{z}_i = \sigma(\mathbf{v}_{EPS,i})$
htobit3	Heteroskedastic tobit model	v_i^*	$\mathbf{x}_i = (v_{NWP,i}, v_{NWP,i}^2, v_{NWP,i}^3)$, $\mathbf{z}_i = \sigma(\mathbf{v}_{EPS,i})$
rq3	Quantile regression	v_i^*	$\mathbf{x}_i = (v_{NWP,i}, v_{NWP,i}^2, v_{NWP,i}^3)$
crq1	Censored quantile regression	v_i^*	$\mathbf{x}_i = v_{NWP,i}$
crq3	Censored quantile regression	v_i^*	$\mathbf{x}_i = (v_{NWP,i}, v_{NWP,i}^2, v_{NWP,i}^3)$
rq3p	Quantile regression in power space	p_i	$\mathbf{x}_i = (f(v_{NWP,i}), f(v_{NWP,i})^2, f(v_{NWP,i})^3)$
srq3p	Quantile regression in power space	p_i	$\mathbf{x}_i =$ 3 spline basis functions of $f(v_{NWP,i})$
srq4wp	Quantile regression in power space	p_i	$\mathbf{x}_i =$ 4 spline basis functions of $v_{NWP,i}$

Table 2.1: List of models considered. The first seven models are all estimated in wind space and all except *rq3* incorporate censoring information. The remaining models are either estimated entirely in power space (*srq3p*, *srq4wp*) or in power-by-wind space (*srq4wp*).

for the data considered. Hence, we confine ourselves to linear functions and order 3 polynomials in $v_{NWP,i}$ for all models in wind space. Only in power space or power-by-wind space, stronger nonlinearities may have to be accounted for by using spline basis functions for (transformed) NWP wind speed. More specifically, we assess three benchmark quantile regression models (*rq3p*, *srq3p*, *srq4wp*) for p_i (i.e., replacing v_i and v_i^* with p_i in Equations 2.14 and 2.15). As regressor variables they either use 3 polynomial basis functions of transformed NWP wind speed forecasts (*rq3p*), spline basis functions (for details see Nielsen et al. 2006) of transformed wind speed forecasts with 3 degrees of freedom (*srq3p*), or spline basis functions of wind speed forecasts with 4 degrees of freedom (*srq4wp*). The benchmark models *srq3p* and *srq4wp* were chosen because they are similar to the models proposed in Nielsen et al. (2006) and Bremnes (2004, 2006), respectively. *srq4wp* is not a local quantile regression model but similar in that it is a nonlinear quantile regression model in the "wind-to-power" space. Model *rq3p* was chosen to investigate differences of spline and polynomial basis functions.

2.4 Verification methodology

In this section, several measures are described to compare the performance of the different models. First a score is introduced in Section 2.4.1 to measure the value of a forecast in a simplified energy market. Such a single value score is very convenient to compare the performance of different forecast methods but unfortunately cannot fully characterize the performance of a forecast (Wilks 2006b). Therefore two important properties of quantile forecasts, reliability and sharpness, are discussed in the following subsections. Reliability is the crucial property of a good forecast that the forecast probabilities match the observed relative

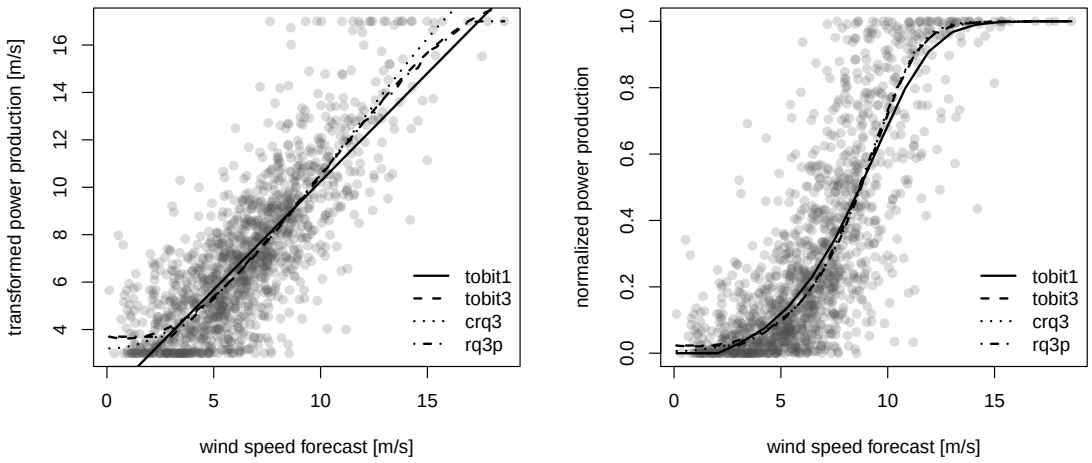

Figure 2.5: Different model fits (median) for the full data set at lead time 24 hours plotted in wind space (left) and power-by-wind space (right).

frequencies. A test to check whether this property is fulfilled is presented in Section 2.4.2. For two reliable forecasts, the one with the narrower predictive distribution is preferable. This property is termed sharpness for which measures are defined in Section 2.4.3.

To evaluate these measures and their variances in an empirical setting, we employ a bootstrapping (Efron and Tibshirani 1994) approach as suggested by Hothorn et al. (2005):

1. Sample n times with replacement from the entire data set (bootstrap sample).
2. Fit the models on this bootstrap sample.
3. Compute performance measures on the "out-of-bootstrap" data, i.e., the observations not contained in the bootstrap sample (approximately 36.8% of the data) .
4. Repeat steps 1–3 k times.

With this approach we obtain $k = 250$ values for each verification measure which can be interpreted as a sample from the associated distribution.

2.4.1 A simple market model score

Since one important application for wind power forecasts is energy trading, the value of a forecast in an energy market can serve as a direct indicator of forecast

performance. Instead of a real energy market, we use a simplified market model (Bremnes 2004; Roulston et al. 2001): First the provider has to bid an amount $\widehat{p}_i$ of energy. The actual production though is p_i. The provider always receives a fee c for the energy p_i he eventually produces. If less than the bid $\widehat{p}_i$ is produced, a penalty c_- for each missing energy unit has to be payed. If too much is produced, each kW of surplus energy is penalized with c_+. Thus, this simple market can be described by the expected income or revenue

$$R(p_i, \widehat{p}_i, c, c_+, c_-) = p_i c - \begin{cases} (\widehat{p}_i - p_i)c_- & \text{if } p_i \leq \widehat{p}_i \\ (p_i - \widehat{p}_i)c_+ & \text{if } p_i > \widehat{p}_i \end{cases} \tag{2.16}$$

In Bremnes (2004) it is shown that the expected income is maximized when $\widehat{p}_i = q_\pi(p_i|\mathbf{x}_i)$, with $\pi = c_+/(c_+ + c_-)$. When dividing Equation 2.16 by $(c_+ + c_-)$, replacing $\widehat{p}_i$ by $q_\pi(p_i|\mathbf{x}_i)$, and using $\pi = c_+/(c_+ + c_-)$ it can be seen that for a specific price combination c, c_- and c_+ the best forecast is the one that minimizes

$$S_{i,\pi} = (1-\pi)(q_\pi(p_i|\mathbf{x}_i) - p_i)I(p_i \leq q_\pi(p_i|\mathbf{x}_i)) + \pi(p_i - q_\pi(p_i|\mathbf{x}_i))I(p_i > q_\pi(p_i|\mathbf{x}_i)) \tag{2.17}$$

Note that this equation is equivalent to the loss function used for quantile regression (Equation 2.13) which is sometimes also referred to as quantile score (Gneiting and Raftery 2007).

A simple performance measure for wind power forecasts would be to compute the income of a specific forecast for a test data set (e.g., as in Bremnes 2004). However, to do so specific market prices have to be assumed. Because prices can vary over different markets and days, we avoid to assume specific market prices by taking the sum of $S_{i,\pi}$ for a range of possible price combinations:

$$S_i = \sum_{j=1}^{9} S_{i,\frac{j}{10}} \tag{2.18}$$

Here, small values of S_i denote good performance. The mean value of S_i over the test dataset is denoted as $\overline{S}$. Note that this score also fits into the framework of Pinson et al. (2007); Gneiting and Raftery (2007) for a unique skill score.

2.4.2 Reliability

Reliability is the property of the forecast probabilities to be in accordance with the observed relative frequencies. For example, 75% of the observations should be on average below the 0.75-quantile. The set of quantile forecasts $q_{1/10}(v_i^*|\mathbf{x}_i), q_{2/10}(v_i^*|\mathbf{x}_i), ..., q_{9/10}(v_i^*|\mathbf{x}_i)$ form 10 intervals with nominal probability of $1/10$ for an observations v_i to fall into one of these intervals. To test the reliability, the relative frequencies of observations falling into specific intervals can be

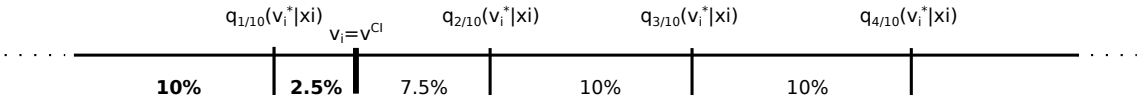

Figure 2.6: Schematic figure how censored observations are split up in a χ^2 test.

compared with their nominal probability by a Pearson's χ^2 test as proposed by Bremnes (2006).

A problem occurs for the censored regression models when the observation falls on one of the censoring points (zero or nominal power). If one or more quantiles are below cut-in or above nominal wind speed respectively it is not clear in which interval the observation falls. Thus, in the χ^2 test such censored observations are split up proportionally into the intervals from which they may stem. To illustrate this split-up strategy, consider the following example (see also Figure 2.6): If the uncensored wind quantiles are $q_{1/10}(v_i^*|\mathbf{x}_i) = 2.5m/s$ and $q_{2/10}(v_i^*|\mathbf{x}_i) = 4.5m/s$ and the observation is censored at cut-in wind speed $v_{CI} = 3$ (i.e, that v_i for which $f^{-1}(p_i = 0)$), then it may either come from the first decile (10%) or the first quarter of the second decile ($2.5\% = (3 - 2.5)/(4.5 - 2.5) * 10\%$). Thus, the first decile receives weight $0.8 = 0.1/(0.1 + 0.025)$ and the second decile receives $0.2 = 0.025/(0.1 + 0.025)$ for this event.

Note that this analysis is done in wind space (before applying Equations 2.2 and 2.3). Exceptions are the models that are estimated in power space for which the analysis is also done in power space. Since censoring is not considered in these models, the split-up of observations is not necessary. Like in Bremnes (2006) we declare forecasts to be unreliable if the p-value of the χ^2 test is below 0.05.

For the quantile regression models, quantile crossing may occur, which makes this reliability analysis difficult. As a simple solution we therefore sort the quantiles before testing. For example if the 0.2 quantile is higher than the 0.3 quantile they are interchanged.

2.4.3 Sharpness

Sharpness is a further property that can be used to characterize forecast performance. Here, we follow the definition of Pinson et al. (2007): Define a central prediction interval as

$$\delta_{\alpha,i} = q_{(1-\alpha/2)}(p_i|\mathbf{x}_i) - q_{\alpha/2}(p_i|\mathbf{x}_i) \tag{2.19}$$

The probability of the observation to fall within this interval is α. Given a reliable forecast, it is preferable that this prediction interval is as narrow as possible, which is related to a small forecast uncertainty. This property is measured by the mean width of the prediction interval over the dataset $\overline{\delta}_\alpha$, which is hereafter

denoted as sharpness.

2.5 Results

In this section the verification measures, introduced in the previous section are used to compare the performance of the different models. Since reliability is the crucial property for a good probabilistic forecast, it is assessed first. Table 2.2 shows the medians of the 250 reliability p-values from bootstrapping of all tested models and lead times. First, it can be seen that all tobit models have worse p-values for lead times 12 and 36 hours than for 24 and 48 hours. While the heteroskedastic tobit model is still reliable for these lead times the p-value of the standard tobit model drops beyond the 0.05 level. A probable reason for this difference can be found in Figure 2.7, which shows the relative frequencies of observations falling into the intervals formed by the predicted deciles from the heteroskedastic tobit model. For both, 12 and 24 hours lead time, the observations fall slightly too often into intervals in the center and too rarely into intervals in the margins (for 36 and 48 hours figures look very similar to 12 and 24 hours respectively and are therefore not shown). This suggests that in fact the response follows a distribution with somewhat heavier tails than the normal distribution. Although this problem is apparent for both lead times it is less pronounced for lead time 24 hours which results in higher reliability p-values than for 12 hours.

When regarding the nonparametric models in Table 2.2, it can be seen that censored quantile regression is reliable for all lead times. A comparison with the uncensored quantile regression shows that not considering the censoring clearly deteriorates the reliability. Finally it can be seen that all models in the power space seem to have problems with reliability whereas the model in the untransformed space (*srq4wp*) is reliable throughout all lead times.

Similar features as in Table 2.2 are shown in Figure 2.8 where a more detailed picture of reliability at lead time 24 hours is plotted. As in Table 2.2, it can be seen that all censored models in wind space (i.e., using the inverse power curve transformation) and the model in the untransformed space are rather reliable while the uncensored quantile regression in wind space (*rq3*) and the models in power space are not.

In Figure 2.9, the market score for different lead times is plotted. Not surprisingly, the market score increases with lead time. Apparently the models predict more poorly for 24 and 48 hours (nighttime) than for 12 and 36 hours. This can be attributed to the fact that in our data set more events with zero production can be found for day time than for night time. For these events mostly a large number of the regarded quantiles are also 0 and therefore S_i small (Equation 2.18).

	12	24	36	48
tobit1	0.02	0.19	0.04	0.21
tobit3	0.03	0.08	0.03	0.18
htobit1	0.08	0.20	0.19	0.30
htobit3	0.08	0.15	0.10	0.20
rq3	0.04	0.06	0.06	0.07
crq1	0.15	0.11	0.14	0.11
crq3	0.19	0.14	0.13	0.11
rq3p	0.00	0.05	0.00	0.06
srq3p	0.00	0.03	0.00	0.05
srq4wp	0.07	0.09	0.09	0.10

Table 2.2: Median p-values (from 250 bootstrap samples) for different lead times (h) from the reliability test for models listed in Table 2.1.

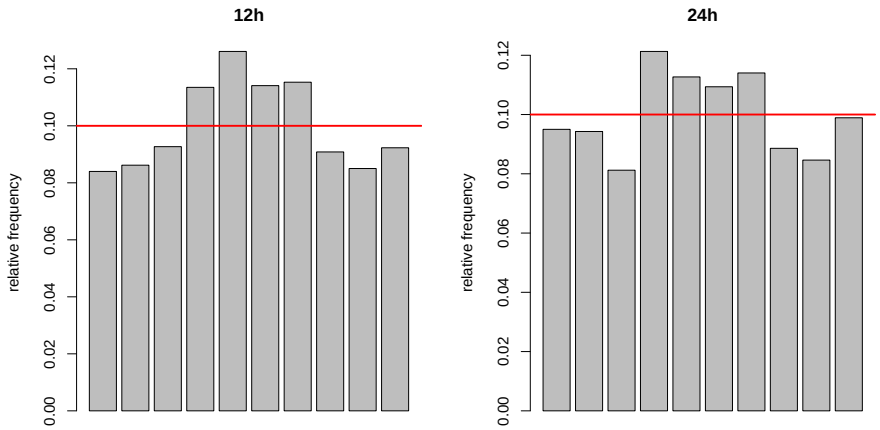

Figure 2.7: Relative frequencies of observations v_i falling into intervals $(-\infty, q_{1/10}(p_i|\mathbf{x}_i)], [q_{1/10}(p_i|\mathbf{x}_i)), q_{2/10}(p_i|\mathbf{x}_i)], \ldots$ for the heteroskedastic tobit model (*htobit3*) for lead time 12 hours (left) and 24 hours (right).

When comparing the models among each other, the differences are small and seem mostly not significant when compared to the uncertainty. To determine if differences are significant, the 250 values from bootstrapping need to be considered pairwise. This can be done, e.g., by using skill scores.

$$SS = 1 - \frac{\overline{S}}{\overline{S}_{REF}} \tag{2.20}$$

where $\overline{S}_{REF}$ is the market score of a reference model which is in our case the quantile regression model with spline basis functions in power space (*srq3p*). Figure 2.10 shows this market skill score for the different models. The heteroskedas-

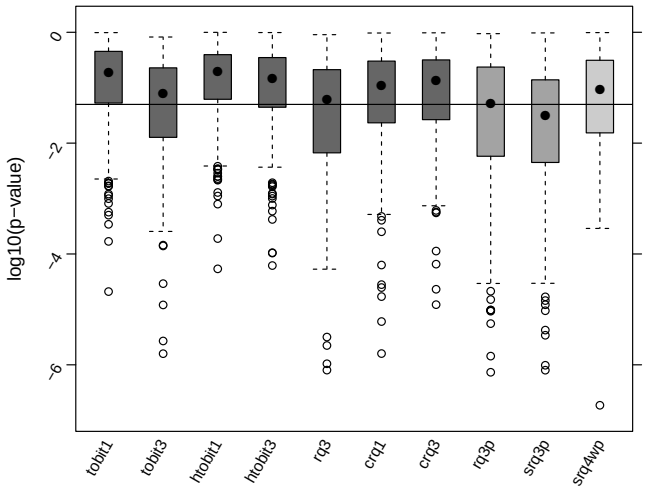

Figure 2.8: Reliability p-values of different models (see Table 2.1) for lead time 24 hours. A horizontal line is plotted for 0.05. The boxes indicate the interquartile ranges of the 250 values from the bootstrapping approach, the whiskers show the most extreme values that are less than 1.5 times the length of the box away from the box, and points are plotted for values that are outside the whiskers.

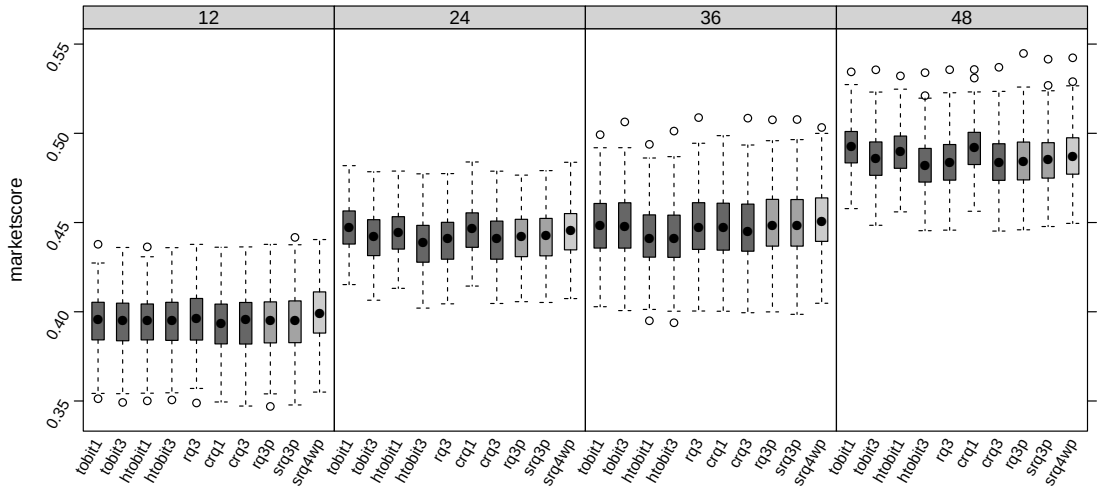

Figure 2.9: Market score ($\overline{S}$; smaller is better) for different models (see Table 2.1) and lead times.

tic tobit model (*htobit3*) performs clearly better than the reference model and the censored quantile regression (*crq3*) is still somewhat better. The remaining models are neither clearly better nor clearly worse than the reference except for the quantile regression model with spline basis functions of wind speed forecasts in the power space (*srq4wp*) which performs worst. Note that the better performance

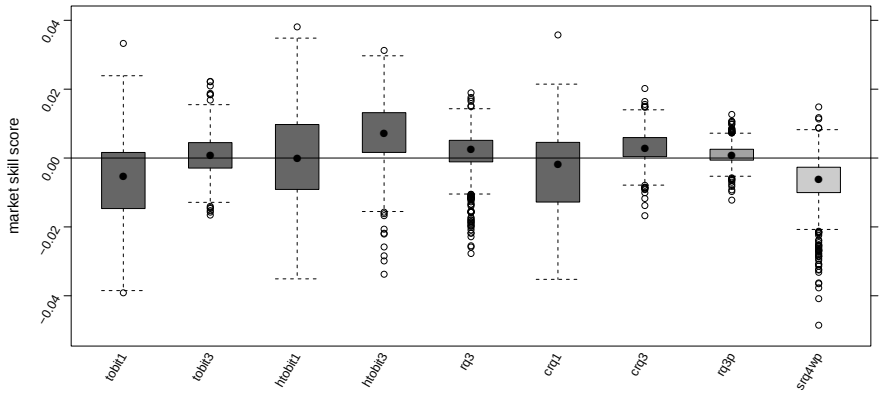

Figure 2.10: Market skill score relative to the reference model *srq3p* (larger is better) for different models (see Table 2.1) and all lead times. Market skill scores greater than 0 indicate better performance than the reference model.

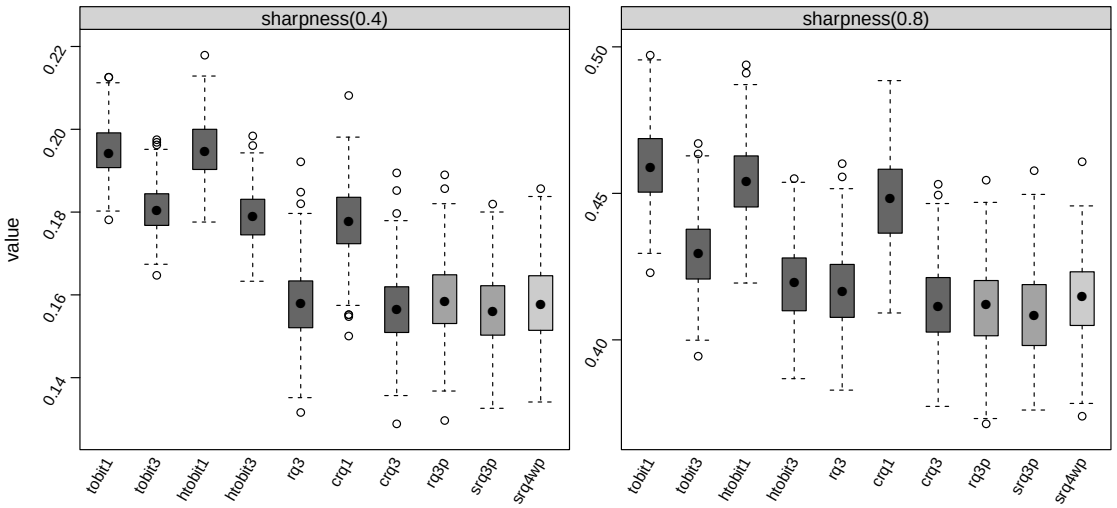

Figure 2.11: Sharpness (smaller is better) of interval forecasts with interval probabilities $\alpha = 0.4$ (left) and $\alpha = 0.8$ (right) for different models (see Table 2.1) and lead time 24 hours.

of the heteroskedastic tobit model (*htobit3*) stems from additional predictive information in form of the ensemble standard deviation of the 10m wind speed ensemble forecast.

The sharpness for two different prediction intervals and lead time 24 hours is shown in Figure 2.11. One feature of this figure is that the nonparametric models have clearly better sharpness, especially for the small 0.4 prediction interval. This can again be attributed to the fact that the assumption of a normal distribution in the parametric models does not apply perfectly (see Figure 2.7).

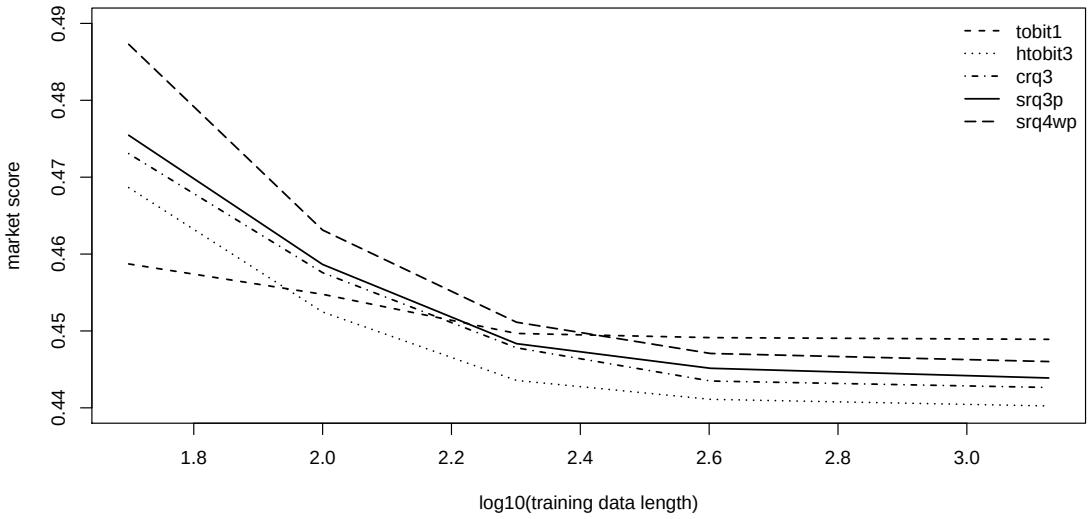

Figure 2.12: Market score ($\overline{S}$; smaller is better) for different training sample sizes, selected models, and lead time 24 hours.

Finally, we show in Figure 2.12 the market scores for different training sample sizes. For computation we used the same bootstrapping approach as described in Section 2.4 but taking smaller samples in step 1. Clearly the performance increases with a larger training sample. Fewer parameters have to be estimated for the parametric models . Therefore it is not surprising that they perform better than the nonparametric models if only few data are available for fitting. The spline model with the completely untransformed data (*rqs4wp*) has the most degrees of freedom and is therefore the worst of all models for small training sample sizes. While for very small training sample sizes the simplest tobit model (*tobit1*) seems to be the best, the heteroskedastic tobit model is already best for training sample sizes ≥ 100. However note that as for the full dataset, the differences are mostly relatively small compared to the uncertainty.

2.6 Conclusion

A combination of new approaches for improving probabilistic wind power forecasts is proposed: (1) Exploit the readily available information from the power curve of the turbine to transform observed power production to wind speed (inverse power curve transformation). (2) Respect the limited range of power production between zero and nominal power (in power space) or between cut-in and nominal wind speed (in wind space) with censored regression models. The resulting combined strategy has the advantage that almost all nonlinearity and

heteroskedasticity of the observations is directly captured. Consequently, relatively simple linear regression models with normally distributed responses can be used.

To assess this new strategy, a wide range of combinations of parametric and nonparametric regression models, with and without inverse power curve transformation, with and without censoring information is considered for data from a wind turbine in Austria. For all models, wind speed forecasts and its transformations are used as regressor variables and, furthermore, some heteroskedasticity models additionally use the standard deviation of the ECMWF ensemble forecasts. It is shown that the censored regression models obtained in wind space with the inverse-transformed power production are more reliable than uncensored regression models in all spaces considered (i.e., in wind space, power space, and power-by-wind space). As for the comparison of parametric vs. nonparametric censored models in wind space, it can be shown that the more parsimonious parametric models already perform well for relatively small training samples while the nonparametric models perform somewhat better in large training samples. However, the performance of the parametric models may potentially be improved in future work by using a response distribution with heavier tails (e.g., logistic or Student-t instead of Gaussian) so that the sharpness can be enhanced.

We have not applied the inverse power curve transformation in combinations with other nonparametric regression models except quantile regression. Nevertheless, the inverse power curve transformation could be applied in combination with any other approach for probabilistic wind power forecasting (e.g., ensemble post processing (Nielsen et al. 2004; Giebel et al. 2005; Pinson and Kariniotakis 2010) or kernel density estimators (Bremnes 2006; Juban et al. 2007; Bessa et al. 2012b,a)). However, consideration of the censoring might be more difficult for these approaches.

In addition to data from the wind turbine presented in this manuscript, data from another wind turbine were assessed but not presented as they lead to very similar outcomes. Hence, similar results can be expected for other turbines/regions but, of course, this still has to be tested in future work. One special feature of the tested turbines is that the wind speeds are relatively small and thus right-censoring (at nominal speed/power) does not play an important role although it is supported by our models. Furthermore, switching off the turbine because of too high wind speed did never happen in our data and is therefore not considered in our models. For turbines where this plays a role it would generally be possible to consider an additional right censoring. Instead of using the manufacturer's power curve, an empirical power curve computed from observation data (Cabezon et al. 2004) could be used as well. This is particularly im-

portant if forecasts for entire wind parks are required which consist of different types of turbines. Our results are based on a global numerical weather prediction (NWP) model rather than a limited area model, as employed by most other studies. However, given the results of Louka et al. (2008); Müller (2011), we do not expect the findings to change much when based on a different NWP model.

Computational details

Our results were obtained on Ubuntu and Debian GNU/Linux using R 2.15.1 (R Core Team 2013) and packages quantreg 4.79 (Koenker 2012) for (censored) quantile regression, and numerical optimization of the likelihood for the (heteroskedastic) tobit models via `optim()` with `method = "BFGS"`. A proper package for the latter is under development but the code is also available upon request in the meantime.

Acknowledgements

This study was supported by the Austrian Science Fund (FWF): L615-N10. The first author was also supported by a PhD scholarship from the University of Innsbruck, *Vizerektorat für Forschung*. We are also very grateful to *WEB Windenergie AG* for providing the wind turbine data. Data from the ECMWF forecasting system were obtained from the ECMWF Data Server. Finally we thank four anonymous reviewers for their comments and suggestions.

3

Paper II

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2013: Heteroscedastic extended logistic regression for post-processing of ensemble guidance. *Monthly Weather Review*, in press.

Heteroscedastic Extended Logistic Regression for Post-Processing of Ensemble Guidance[1]

Jakob W. Messner[2] and Georg J. Mayr
Institute of Meteorology and Geophysics, University of Innsbruck, Austria

Daniel S. Wilks
Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

Achim Zeileis
Department of Statistics, University of Innsbruck, Austria

ABSTRACT

To achieve well-calibrated probabilistic forecasts, ensemble forecasts are often statistically post-processed. One recent ensemble-calibration method is extended logistic regression which extends the popular logistic regression to yield full probability distribution forecasts. Although the purpose of this method is to post-process ensemble forecasts, usually only the ensemble mean is used as predictor variable, whereas the ensemble spread is neglected because it does not improve the forecasts. In this study we show that when simply used as ordinary predictor variable in extended logistic regression, the ensemble spread only affects the location but not the variance of the predictive distribution. Uncertainty information contained in the ensemble spread is therefore not utilized appropriately. To solve this drawback we propose a new approach where the ensemble spread is directly used to predict the dispersion of the predictive distribution. With wind speed data and ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) we show that using this approach, the ensemble spread can be used effectively to improve forecasts from extended logistic regression.

3.1 Introduction

Weather forecasts are very important for many parts of social and economic life. For example they are used for severe weather warnings, for decision making in

[1]in press in *Monthly Weather Review*

[2]*Corresponding author address:* Institute of Meteorology and Geophysics, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria. E-mail: jakob.messner@uibk.ac.at

agriculture and industry, or for planning of leisure activities. Generally these forecasts are based on numerical weather prediction (NWP) models. Unfortunately, because of uncertainties in the initial conditions and unknown or unresolved atmospheric processes these models are always subject to error. Luckily some of these errors are systematic and can be corrected with statistical post-processing, often also referred to as model output statistics (MOS; Glahn and Lowry (1972)). However, not all errors can be corrected and for many customers it is important to get additional information about the remaining forecast uncertainty. For this purpose many forecasting centers provide ensemble forecasts. These are multiple NWP forecasts with slightly perturbed initial conditions and sometimes also different model formulations. The idea is that these different forecasts should represent the range of possible outcomes (Lorenz 1996). Large ensemble spreads are then presumably associated with high forecast uncertainties and small spreads signify low uncertainties. However, in practice the initial ensemble members do not represent initial-condition uncertainty (Hamill et al. 2003; Wang and Bishop 2003). Furthermore ensemble forecasts exhibit the same model errors as single integration forecasts. Thus, to achieve unbiased and calibrated uncertainty forecasts, statistical post-processing is needed.

In the past decade much research has gone into finding appropriate methods to post-process ensemble forecasts. For example Roulston and Smith (2003) proposed dressing the ensemble members with historical model errors and Raftery et al. (2005) suggested Bayesian model averaging for this purpose. Gneiting et al. (2005) proposed to use linear regression with error variances depending on the ensemble spread, and for binary predictands Hamill et al. (2004) proposed to use logistic regression. Comparisons of these and other methods (Wilks 2006a; Wilks and Hamill 2007) showed that logistic regression is one of the better approaches. A very promising extension of logistic regression has been proposed recently (Wilks 2009). By including the predictand threshold in the regression equations this extended logistic regression allows derivation of full predictive distributions. The extended logistic regression method has been used in several studies for probabilistic precipitation forecasts (Schmeits and Kok 2010; Ruiz and Saulo 2012; Roulin and Vannitsem 2012; Hamill 2012; Ben Bouallègue 2013; Scheuerer 2013) and was shown to perform very well compared to standard logistic regression (Wilks 2009; Ruiz and Saulo 2012) and other ensemble post-processing methods (Schmeits and Kok 2010; Ruiz and Saulo 2012; Scheuerer 2013). In all of these studies, extended logistic regression is used to post-process ensemble forecasts, but usually the ensemble mean was used as the only predictor variable. There were also several attempts to additionally include the ensemble spread, but with the exception of Hamill (2012) it was always disregarded because it did not im-

prove the forecasts.

In this study we show that the predictive distribution of the transformed predictand is logistic and that the predictor variables only affect the location (mean) but not the dispersion (variance) of this logistic distribution. So far the ensemble spread was always included as ordinary predictor variable in extended logistic regression so that its information was only used to predict the location but not the dispersion of the forecast distribution. However, the ensemble spread is generally expected to mainly contain information about the forecast uncertainty which in turn should be directly related to the dispersion of the predictive distribution. Hence, the uncertainty information contained in the ensemble spread cannot be utilized properly by extended logistic regression so that it is not surprising that no improvements could be found.

To solve this drawback of extended logistic regression, we therefore propose a simple new approach in which the ensemble spread can be directly used as predictor for the dispersion of the forecast probability distribution. To illustrate our findings and test if improvements can be achieved with this new approach, we compare different approaches to include the ensemble spread in extended logistic regression on wind speed data from 11 European locations and ensemble forecasts from the European Centre for Medium Range Weather Forecasts (ECMWF).

The remainder of the paper is organized as follows: In Section 3.2 we describe the extended logistic regression model and show the problems when including the ensemble spread as ordinary predictor variable. Our new approach is introduced in Section 3.3. Results from the case study are shown in Section 3.4 and a summary and conclusion can be found in Section 3.5.

3.2 Extended logistic regression

Originally, logistic regression is a regression model from the generalized linear model framework (Nelder and Wedderburn 1972) to model the conditional probability of binary events. As such it is also a well-suited MOS method for binary predictands (Hamill et al. 2004). For example, the probability of a continuous variable y to fall below a certain threshold q can be predicted with:

$$P(y < q|\mathbf{x}) = \frac{\exp(\mathbf{x}^\top \beta)}{1 + \exp(\mathbf{x}^\top \beta)} = \Lambda(\mathbf{x}^\top \beta) \tag{3.1}$$

where $\mathbf{x}$ is a vector of predictor variables (e.g., NWP forecasts; $\mathbf{x} = (1, x_1, x_2, ...)^\top$) and β is a vector of regression coefficients ($\beta = (\beta_0, \beta_1, \beta_2, ...)^\top$) that is generally estimated with maximum likelihood estimation (see Appendix). The regression function has the same mathematical form as the cumulative distribution function

of the standard logistic distribution (Λ) which is indicated by the final equality in Equation 3.1.

Often, more than one threshold is of interest and separate logistic regressions are fitted for each of these thresholds. This approach has the disadvantage that the predicted probabilities are not constrained to be mutually consistent. In other words, for two thresholds q_a and q_b with $q_a < q_b$ it can occur that $P(y < q_a|\mathbf{x}) > P(y < q_b|\mathbf{x})$ which would imply nonsense negative probabilities for $P(q_a \leq y < q_b|\mathbf{x})$.

To avoid these inconsistencies, Wilks (2009) extended logistic regression by including (a transformation of) the thresholds q_j as additional predictor variable.

$$P(y < q_j|\mathbf{x}) = \Lambda(\alpha g(q_j) + \mathbf{x}^\top \beta) \tag{3.2}$$

Here $g(q_j)$ is a nondecreasing function of q_j and α is an additional coefficient that has to be estimated. In addition to avoiding negative probabilities, this extended logistic regression has the advantage that fewer coefficients have to be estimated (instead of different vectors β for each threshold, α and β are the same for all thresholds), which is especially advantageous for small training data sets (Wilks 2009). Furthermore, the probability to fall below any arbitrary value Q can be easily computed by replacing q_j with Q:

$$P(y < Q|\mathbf{x}) = \Lambda(\alpha g(Q) + \mathbf{x}^\top \beta) \tag{3.3}$$

Equation 3.3 can also be interpreted as continuous cumulative distribution function which implies that full continuous probability distributions can be provided.

Since $g()$ has to be a nondecreasing function, the equation

$$P(y < Q|\mathbf{x}) = P(g(y) < g(Q)|\mathbf{x}) \tag{3.4}$$

is always fulfilled. With Equation 3.4 and some rearrangements, Equation 3.3 can also be written as

$$P(g(y) < g(Q)|\mathbf{x}) = \Lambda\left(\frac{g(Q) + \mathbf{x}^\top \beta/\alpha}{1/\alpha}\right) \tag{3.5}$$

and upon setting $\mu = -\mathbf{x}^\top \beta/\alpha$ and $\sigma = 1/\alpha$ we obtain

$$P(g(y) < g(Q)|\mathbf{x}) = \Lambda\left(\frac{g(Q) - \mu}{\sigma}\right) \tag{3.6}$$

This notation allows one to easily see that the conditional probability distribution of the transformed predictand $g(y)$ given the predictor variables x is a logistic distribution with location parameter μ and scale parameter σ. Cumulative distribution functions and probability density functions of this distribution with different

scale parameters σ are shown in Figure 3.1. The shape of the logistic distribution is very similar to that of the normal distribution but with somewhat heavier tails. The mean of this distribution is μ and in terms of the scale parameter the variance is $\sigma^2\pi^2/3$ (Johnson et al. 1995).

Note that the scale parameter $\sigma = 1/\alpha$ is constant so that the predictor variables in $\mathbf{x}$ only affect the mean but not the variance of the logistic predictive distribution. Hence, when included as additional predictor variable in $\mathbf{x}$, the ensemble spread has no effect on the dispersion of the predictive distribution. However, usually large ensemble spreads are associated with high forecast uncertainties, which in turn should be related to wider predictive distributions. In contrast the level of uncertainty should generally have no effect on the location of the forecast probability distribution.

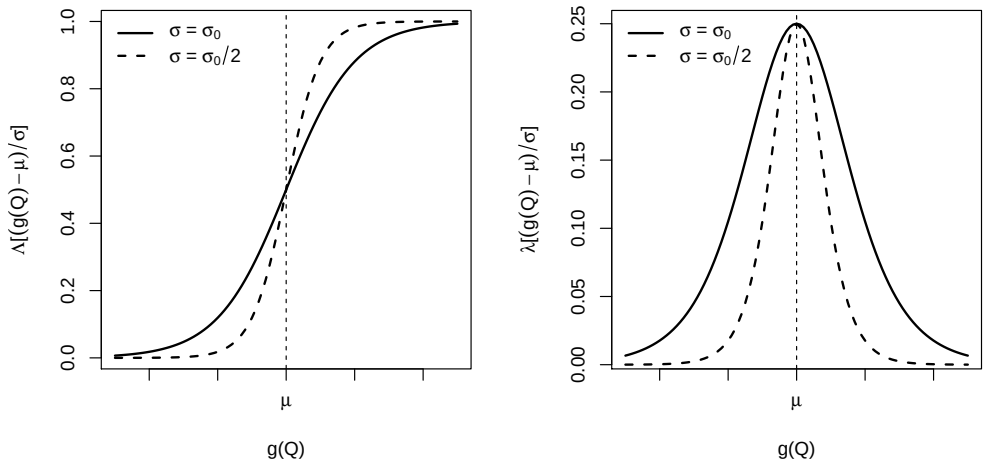

Figure 3.1: Cumulative distribution function (left) and probability density function (right) of the logistic distribution for different scale parameters σ. Here $\lambda(x) = \frac{d\Lambda(x)}{dx}$ is the probability density function of the standard logistic distribution.

3.3 Heteroscedastic extended logistic regression

In the previous section we showed that when using the ensemble spread as an ordinary predictor variable in extended logistic regression, uncertainty information is not utilized appropriately. As a more effective approach we therefore propose to use the ensemble spread directly as predictor for the *dispersion* of the predictive

distribution. Therefore we simply replace μ and σ in Equation 3.6 with

$$\mu = \mathbf{x}^\top \gamma \tag{3.7}$$

and

$$\sigma = \exp(\mathbf{z}^\top \delta) \tag{3.8}$$

respectively. Here $\mathbf{z}$ is an additional vector of input variables (i.e., the ensemble spread) and γ and δ are coefficient vectors that have to be estimated. The exponential function is used here as a simple method to ensure positive values for σ.

Note that with $\mathbf{z} = 1$ this model is completely equivalent to the original extended logistic regression (Equation 3.2) with $\alpha = 1/\exp(\delta)$ and $\beta = -\gamma/\exp(\delta)$.

The idea of using the ensemble spread as predictor for the dispersion is not completely new. For Gaussian linear regression models, Gneiting et al. (2005) proposed a similar approach, which has been proven to perform well in several studies (e.g., Wilks 2006a; Wilks and Hamill 2007).

3.4 Case study

In this section, we apply the findings from the previous sections on real data. We use 10 meter wind speed observations (mean over last 10 minutes) from the 11 European weather stations *Amsterdam–Schiphol* (52.3 N, 4.783 E), *Berlin–Tegel* (52.55 N, 13.3 E), *Brussels–National* (50.9 N, 4.533 E), *Copenhagen–Airport* (55.6 N, 12.633 E), *Frankfurt–Main* (50.033 N, 8.583 E), *London–Heathrow* (51.467 N, −0.45 E), *Lisbon–Geof* (38.767 N, −9.133 E), *Madrid–Barajas* (40.467 N, −3.55 E), *Paris–Orly* (48.717 N, 2.383 E), *Rome–Fiumicino* (41.8 N, 12.233 E), and *Wien–Hohe-Warte* (48.249 N, 16.356 E), from April 2010 to December 2012. As NWP forecasts we use ensemble wind speed forecasts bilinearly interpolated to the instrument location from the European Centre for Medium Range Weather Forecasts (ECMWF; Molteni et al. 1996), initialized at 00 UTC for the lead times 24, 36, 48, and 60 hours.

Figure 3.2 shows a clear positive correlation between ensemble spread and forecast error for *Wien–Hohe-Warte* (similar for most other locations). This positive spread-skill relationship suggests that the ensemble spread contains potentially useful uncertainty information. To investigate how this information might be used most effectively, we compare different extended logistic regression models.

For all models we use the square root function for $g(\cdot) = \sqrt{\cdot}$. This function gave good results for precipitation forecasts in several studies (Wilks 2009; Schmeits and Kok 2010; Roulin and Vannitsem 2012; Ruiz and Saulo 2012;

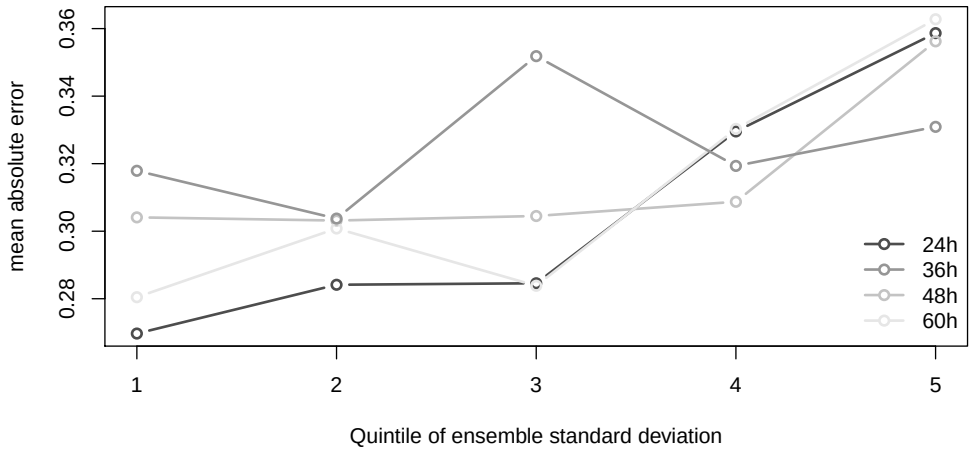

Figure 3.2: Mean absolute error of ensemble median for different ensemble standard deviations and lead times computed for *Wien–Hohe-Warte* (33 months of data). Quintiles are used to divide the ensemble standard deviation into different levels. Note that for this plot all wind speeds are square-root transformed.

Model		$\mathbf{x}$	$\mathbf{z}$
XLR	Extended logistic regression	$(1, M)^\top$	1
XLR:S	Extended logistic regression	$(1, M, S)^\top$	1
XLR:SM	Extended logistic regression	$(1, M, S * M)^\top$	1
HXLR	Heteroscedastic extended log. reg.	$(1, M)^\top$,	$(1, S)^\top$
HXLR:S	Heteroscedastic extended log. reg.	$(1, M, S)^\top$,	$(1, S)^\top$

Table 3.1: List of different extended logistic regression models. $\mathbf{x}$ and $\mathbf{z}$ are vectors of predictor variables for the location and scale of the predictive distribution respectively. M and S are the mean and standard deviation of square root transformed wind speed ensemble forecasts respectively.

Ben Bouallègue 2013; Scheuerer 2013) and also improves our wind speed forecasts compared to the identity function. As potential regressors we use the ensemble mean (M) and standard deviation (S) of the square-root-transformed ensemble wind speed forecasts. Furthermore we selected $J = 9$ climatological quantiles with probabilities $1/10$, $2/10$, ... $9/10$ as thresholds q_j for each location separately.

Table 3.1 lists the models that are used in the following. In addition to the extended logistic regression model with the ensemble mean as single predictor variable (*XLR*) there are 4 models which use the ensemble standard devia-

tion. The models *XLR:S* and *XLR:SM* are standard extended logistic regression models with the ensemble standard deviation as additional predictor variable, either alone (*XLR:S*) or multiplied with the ensemble mean (*XLR:SM*). In the heteroscedastic extended logistic regression model *HXLR* the ensemble standard deviation is only included as predictor variable for the scale and in *HXLR:S* it is additionally also used as predictor variable for the location of the predictive distribution.

Before reporting the forecast quality of these different models it is interesting to investigate the effect of the ensemble spread on the predicted probability distributions. Figure 3.3 shows predicted probability density functions of the *XLR:S* and *HXLR* models for different ensemble standard deviations. For the *XLR:S* model it can be seen that contrary to the desired effect, larger ensemble standard deviations are related to slightly sharper distributions. In contrast, the *HXLR* model uses the ensemble standard deviation more appropriately and larger ensemble standard deviations are clearly related to wider distributions.

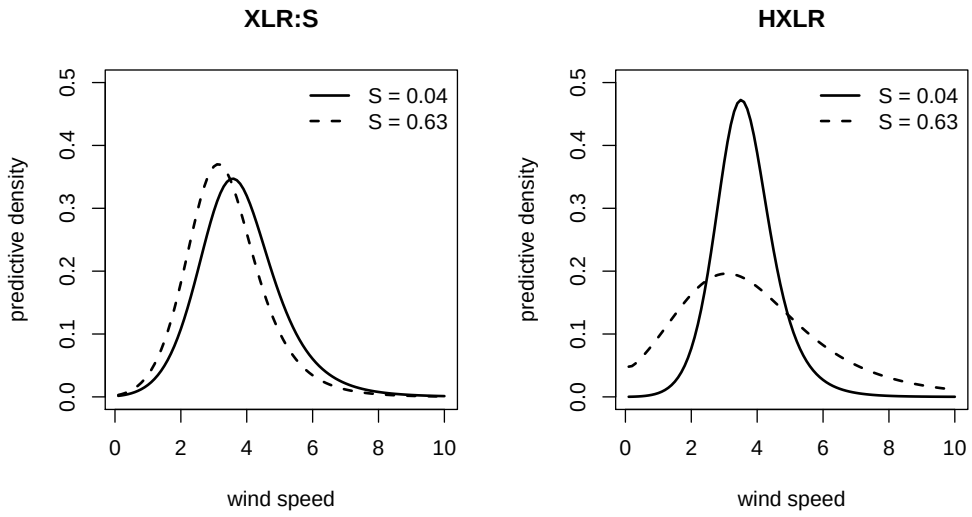

Figure 3.3: Predicted probability density functions of *XLR:S* (left) and *HXLR* (right) for small and large ensemble standard deviations respectively. The models (see Table 3.1 for details) are fitted for *Wien–Hohe-Warte* and 36 hours lead time. For all curves the ensemble mean is $M = 2$, which is approximately the mean ensemble mean of the data set. The ensemble standard deviations 0.04 and 0.63 are approximately the minimum and maximum ensemble standard deviation in the data set respectively.

Next we compare the performance of the different models. Since extended logistic regression can provide multi-categorical probabilistic forecasts, the ranked

probability score (Epstein 1969; Wilks 2006b) is a well-suited measure of forecast quality.

$$RPS = \sum_{j=1}^{J} (P(y < q_j|\mathbf{x}) - I(y < q_j))^2 \tag{3.9}$$

Here $J = 9$ is the number of thresholds and $I(\cdot) = 1$ if the argument in brackets is true and 0 if it is not. To get independent training and test data sets we estimate and verify the models with 10-fold cross validation. With this cross validation we get one *RPS* value for each event in the dataset. From these individual *RPS*, 250 estimates of the mean ($\overline{RPS}$) are computed on 250 bootstrap samples. This is all done separately for each model, location, and lead time. Since we are mainly interested in improvements that can be achieved with the ensemble standard deviation we finally compute skill scores (*RPSS*) with the standard extended logistic regression model (*XLR*) as reference.

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}_{XLR}} \tag{3.10}$$

Note that here positive values indicate improvements over the standard extended logistic regression model.

Figure 3.4 shows the *RPSS* of the different models and lead times aggregated over the 11 locations. It can be seen that including the ensemble standard deviation simply as ordinary predictor variable (*XLR:S*, *XLR:SM*) does not improve forecast quality of extended logistic regression. However, the reason is not the absence of predictive information in the ensemble standard deviation since using it with our new approach (*HXLR*) clearly improves the forecast quality, especially for day time forecasts (36 and 60 hours lead time). Since the ensemble standard deviation seems not to contain any predictive information on the location it is also not advantageous to include it additionally as predictor variable for the location (*HXLR:S*). The effect of the lead time on the *RPSS* is only weak but for day time forecasts ($12h$, $36h$) the superiority of *HXLR* is more pronounced. Note that we also tested longer lead times (up to 96 hours) and shorter training data lengths (down to 6 months) but results were similar and are therefore not shown.

Figure 3.5 shows the *RPSS* for selected locations aggregated over lead times 24 to 60 hours. While most of the locations show similar patterns as in Figure 3.4 (e.g. Amsterdam, Wien) there are also some locations (e.g. Berlin) where including the ensemble spread as ordinary predictor variable (*XLR:S*, *XLR:SM*) is superior to heteroscedastic extended logistic regression (*HXLR*). This suggests that for these locations the ensemble spread also contains predictive information on the location. For non-negative predictands like wind speed, large observed values are generally related to large ensemble spreads. Therefore it is indeed

conceivable that the ensemble spread contains some predictive information on the location that is not yet covered by the ensemble mean. However, additional improvements can be achieved when including the ensemble spread as predictor for both, location and scale of the predictive distribution (*HXLR:S*).

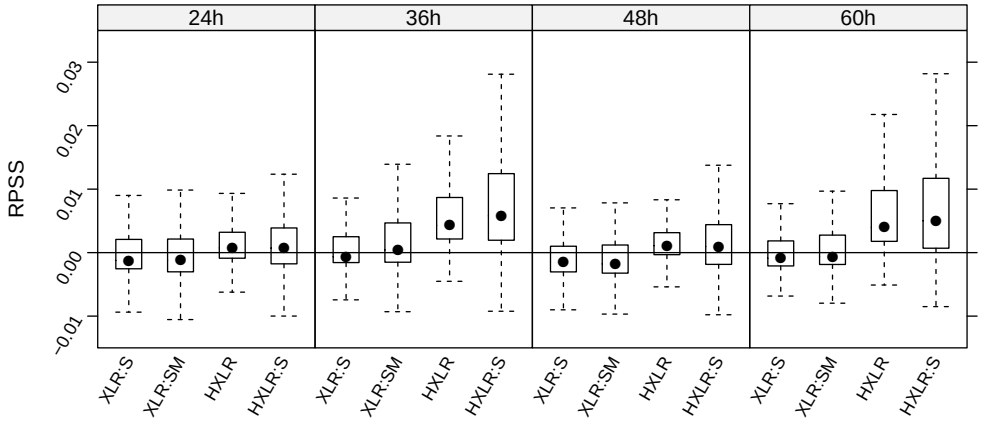

Figure 3.4: Ranked probability skill score (*RPSS*) relative to extended logistic regression (*XLR*) for different lead times and models (see Table 3.1 for details) aggregated over 11 European locations (see text for details). 9 climatological deciles are used as thresholds. Positive values indicate improvements over *XLR*. The boxes indicate the interquartile ranges of the 11*250 values from the bootstrapping approach and the whiskers show the most extreme values that are less than 1.5 times the length of the box away from the box (further outliers have been omitted).

Finally, Figure 3.6 shows reliability diagrams for 36 hour forecasts of the first climatological decile ($P(y < q_1|\mathbf{x})$) and the climatological median $P(y < q_5|\mathbf{x})$ for the models *XLR:S* and *HXLR*. Both models are fairly reliable with only little differences between each other. For the lower decile both models are slightly over-forecasting (points below diagonal). The logistic predictive distribution of extended logistic regression involves a point mass at zero (i.e. positive predictive density for negative wind speeds; Schefzik et al. 2013). Because zero wind speeds occur relatively rarely, this might be the reason for the overestimated probabilities to fall below the lower decile.

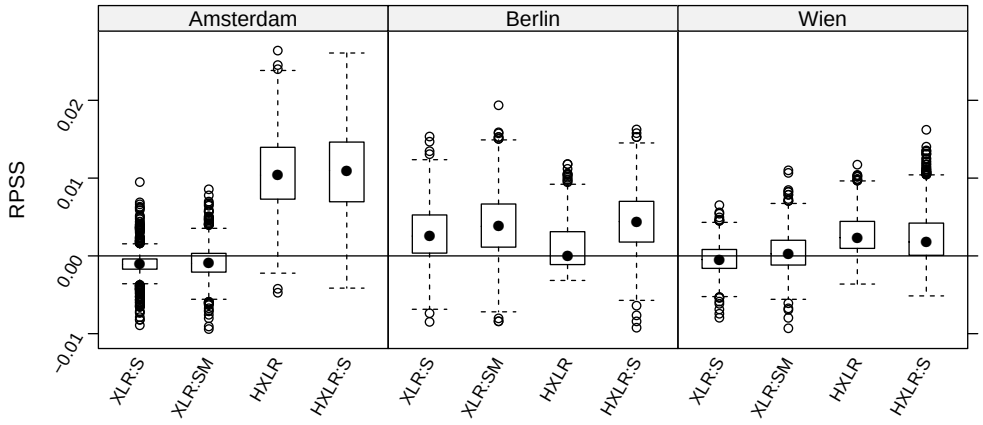

Figure 3.5: Ranked probability skill score (*RPSS*) relative to extended logistic regression (*XLR*) for selected locations, aggregated over lead times 24, 36, 48, and 60 hours. 9 climatological deciles are used as thresholds. The boxes indicate the interquartile ranges of the 4*250 values from the bootstrapping approach and the whiskers show the most extreme values that are less than 1.5 times the length of the box away from the box. Further outliers are plotted as circles

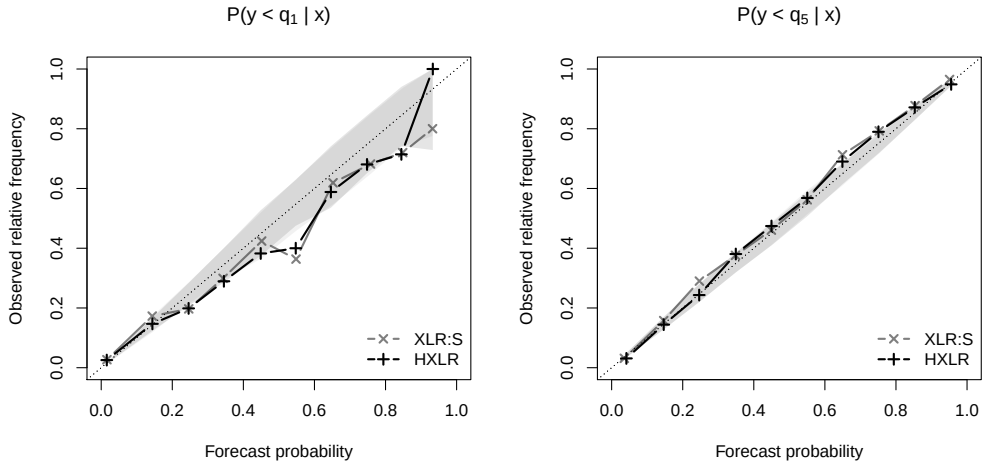

Figure 3.6: Reliability diagrams for the 36 hour probability forecasts $P(y < q_1|\mathbf{x})$ (left; 1st climatological decile) and $P(y < q_5|\mathbf{x})$ (right; climatological median) from *XLR:S* (gray) and *HXLR* (black). Forecasts are pooled for all locations and aggregated in 0.1 intervals. The gray areas show the 95% consistency intervals for *XLR:S* and *HXLR* (with alpha-blending) derived from consistency re-sampling (Bröcker and Smith 2007)

3.5 Summary and conclusion

The inclusion of the ensemble spread in extended logistic regression has been shown in several studies not to improve the forecast skill. As we have shown in this paper this is not surprising because when the ensemble spread is included as ordinary predictor variable it modifies only the location but not the dispersion of the forecast distribution. Uncertainty information contained in the ensemble spread is therefore not used appropriately. To solve this problem we proposed a new approach called heteroscedastic extended logistic regression where the ensemble spread is directly used as predictor for the *scale* of the predictive distribution.

To illustrate the advantages of this new approach we used wind speed observations from 11 European locations and ensemble forecasts from ECMWF. Consistent with our findings and results from previous studies, the inclusion of the ensemble standard deviation as an ordinary predictor variable has no clear positive effects on forecast quality. In contrast, with our new approach the uncertainty information in the ensemble standard deviation is used effectively to achieve clear improvements.

An additional single case study with precipitation data showed similar results. We therefore expect that our results can be transferred to other variables and/or locations. However, this still has to be tested.

Hamill (2012) got better forecasts when using the ensemble variance multiplied with the ensemble mean as additional predictor variable. This suggests that in his data the ensemble spread also contained predictive information on the location of the predictive distribution. Consistent with these findings, we also found individual weather stations where including the ensemble spread as ordinary predictor variable is even superior to heteroscedastic extended logistic regression. However, further improvements could be achieved when including the ensemble spread as predictor variable for both, location and spread of the predictive distribution.

To enhance the flexibility of extended logistic regression, Ben Bouallègue (2013) proposed to use interaction terms between the threshold and the predictor variables. An interaction term between threshold and ensemble spread could also be used to control the dispersion of the predictive distribution. Contrary to heteroscedastic extended logistic regression such a model can be easily implemented with standard binary logistic regression software. However, with interaction terms the ensemble spread also has some undesired effects on the distribution location.

Extended logistic regression has been shown in several studies to perform

well compared to other ensemble post-processing algorithms (e.g., Schmeits and Kok 2010; Ruiz and Saulo 2012; Scheuerer 2013). However, a major drawback of this method was that uncertainty information contained in the ensemble spread could not be utilized effectively. Heteroscedastic extended logistic regression is therefore a very attractive extension of extended logistic regression to further enhance its competitiveness.

Computational details

Our results were obtained on Ubuntu using R 2.15.2 (R Core Team 2013). The (heteroscedastic) extended logistic regression models were built upon the package ordinal 2012.09-11 (Christensen 2013). A proper package is under development but the code is also available upon request in the meantime.

Acknowledgements

We thank Tom Hamill, Tilmann Gneiting, Constantin Junk, and an anonymous reviewer for their valueable comments to improve this manuscript. This study was supported by the Austrian Science Fund (FWF): L615-N10. The first author was also supported by a PhD scholarship from the University of Innsbruck, *Vizerektorat für Forschung*. Data from the ECMWF forecasting system were obtained from the ECMWF Data Server.

A: Likelihood function

To estimate the coefficients α and β (extended logistic regression) or γ and δ (heteroscedastic extended logistic regression) maximum likelihood estimation is used. The general log-likelihood function for logistic regression models is

$$l = \sum_{i=1}^{N} \log(\pi_i) \tag{3.11}$$

where N is the length of the training data set and π_i is the predicted probability for the i-th observed outcome. For binary logistic regression there are two possible outcomes, so that

$$\pi_i = \begin{cases} P(y_i < q|\mathbf{x}_i) & y_i < q \\ 1 - P(y_i < q|\mathbf{x}_i) & y_i \geq q \end{cases} \tag{3.12}$$

In previous studies the sum of this binary log-likelihood over all thresholds is used as objective function that is maximized to estimate the regression coefficients. However, the predicted probability of the i-th outcome actually is

$$\pi_i = \begin{cases} P(y_i < q_1|\mathbf{x}_i) & y_i < q_1 \\ P(y_i < q_j|\mathbf{x}_i) - P(y_i < q_{j-1}|\mathbf{x}_i) & q_{j-1} \leq y_i < q_j \\ 1 - P(y_i < q_M|\mathbf{x}_i) & y_i \geq q_M \end{cases} \tag{3.13}$$

so that the correct maximum likelihood estimator is given by the maximization of Equations 3.11 and 3.13. In this study we employ this maximum likelihood estimator to take advantage of all standard asymptotic inference in the maximum likelihood framework. However, the concepts presented in this paper do not depend on the objective function and results should also not differ significantly when using the sum of binary log-likelihoods (Equation 3.12) to estimate the coefficients.

4

Paper III

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2013: Extending extended logistic regression for ensemble post-processing: extended vs. separate vs. ordered vs. censored. *Monthly Weather Review*, submitted.

Extending Extended Logistic Regression for Ensemble Post-Processing:
Extended vs. Separate vs. Ordered vs. Censored[1]

Jakob W. Messner[2] and Georg J. Mayr
Institute of Meteorology and Geophysics, University of Innsbruck, Austria

Daniel S. Wilks
Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

Achim Zeileis
Department of Statistics, University of Innsbruck, Austria

ABSTRACT

Extended logistic regression is a recent ensemble calibration method that extends logistic regression to provide full continuous probability distribution forecasts. It assumes conditional logistic distributions for the (transformed) predictand and fits these using selected predictand category probabilities. In this study we compare extended logistic regression to the closely related ordered and censored logistic regression models. Ordered logistic regression avoids the logistic distribution assumption but does not yield full probability distribution forecasts, whereas censored regression directly fits the full conditional predictive distributions.

To compare the performance of these and other ensemble post-processing methods we used wind speed and precipitation data from two European locations and ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). Ordered logistic regression performed similarly to extended logistic regression for probability forecasts of discrete categories whereas full predictive distributions were better predicted by censored regression.

[1]submitted to *Monthly Weather Review*

[2]*Corresponding author address:* Institute of Meteorology and Geophysics, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria. E-mail: jakob.messner@uibk.ac.at

4.1 Introduction

Important applications such as severe weather warnings or decision making in agriculture, industry, and finance strongly demand accurate weather forecasts. Usually numerical weather prediction (NWP) models are used to provide these weather forecasts. Unfortunately, because of the only roughly known current state of the atmosphere and unknown or unresolved physical processes, NWP models are always subject to error. To estimate these errors many forecasting centers nowadays provide ensemble forecasts. These are several NWP forecasts with perturbed initial conditions and/or different model formulations. However, the perturbed initial conditions do not necessarily represent initial condition uncertainty (Hamill et al. 2003; Wang and Bishop 2003) and some structural deficiencies in the models are also not accounted for. Thus, the ensemble forecasts usually do not represent the full uncertainty of NWP models. Ensemble forecasts therefore typically need to be statistically post-processed to achieve well-calibrated probabilistic forecasts.

In the past decade a variety of different ensemble post-processing methods have been proposed. Examples are ensemble dressing (Roulston and Smith 2003), Bayesian model averaging (Raftery et al. 2005), heteroscedastic linear regression (Gneiting et al. 2005), or logistic regression (Hamill et al. 2004). Comparisons of these and other post-processing methods (Wilks 2006a; Wilks and Hamill 2007) showed that logistic regression performs relatively well. Recently, Wilks (2009) extended logistic regression by including the (transformed) predictand thresholds as an additional predictor variable. In addition to requiring fewer coefficients and providing coherent probabilistic forecasts this extended logistic regression allows derivation of full continuous predictive distributions. Extended logistic regression has been used frequently (Schmeits and Kok 2010; Ruiz and Saulo 2012; Roulin and Vannitsem 2012; Hamill 2012; Ben Bouallègue 2013; Scheuerer 2013; Messner et al. 2013c) and has been further extended to additionally account for conditional heteroscedasticy (Messner et al. 2013c). Recently, several studies noticed that extended logistic regression assumes a conditional logistic distribution for the transformed predictand (Scheuerer 2013; Schefzik et al. 2013; Messner et al. 2013c) where this logistic distribution is fitted to selected predictand category probabilities.

In this study we compare (heteroscedastic) extended logistic regression with two closely related regression models from statistics that are particularly popular in econometrics (and more broadly the social sciences):

1. (Heteroscedastic) ordered logistic regression also provides coherent forecasts of category probabilities. However it differs from extended logistic

regression in that no continuous distribution is assumed or specified by the model.

2. (Heteroscedastic) censored regression also fits conditional logistic distributions to a transformed predictand but employs the full set of training-data points (as opposed to a set of thresholds) for fitting the model.

The performance of these statistical models is tested on wind speed and precipitation data from two European locations and ensemble forecasts from the European Centre for Medium Range Weather Forecasts (ECMWF). In addition to heteroscedastic ordered logistic regression, heteroscedastic extended logistic regression, and heteroscedastic censored logistic regression, also separate logistic regressions (Hamill et al. 2004) and for wind speed forecasts heteroscedastic truncated Gaussian regression (Thorarinsdottir and Gneiting 2010) are tested.

The following Section 4.2 describes the different statistical models in detail. A brief description of the data can be found in Section 4.3. Finally, Section 4.4 presents the results that are summarized and discussed in Section 4.5.

4.2 Statistical models

This section describes different statistical models to predict conditional probabilities $P(y \leq q_j|\mathbf{x})$ of a continuous predictand y falling below a threshold q_j, given a vector of predictor variables $\mathbf{x} = (1, x_1, x_2, \dots)^\top$ (i.e., NWP forecasts). Conditional category probabilities of y to fall between two thresholds q_a and q_b can then easily be derived with $P(q_a < y \leq q_b) = P(y \leq q_b|\mathbf{x}) - P(y \leq q_a|\mathbf{x})$.

4.2.1 Separate logistic regressions (SLR)

Logistic regression was one of the first statistical methods that were proposed to post-process ensemble forecasts (Hamill et al. 2004). Originally it is a regression model from the generalized linear model framework (Nelder and Wedderburn 1972) to model the probability of binary responses:

$$P(y \leq q_j|\mathbf{x}) = \frac{\exp(\mathbf{x}^\top \beta)}{1 + \exp(\mathbf{x}^\top \beta)} = \Lambda(\mathbf{x}^\top \beta) \tag{4.1}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \dots)^\top$ is a coefficient vector and $\Lambda(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$ is notationally equivalent to the cumulative distribution function of the standard logistic distribution. The coefficient vector β is estimated by maximizing the the log-likelihood

$$\ell = \sum_{i=1}^{N} \log(\pi_i) \tag{4.2}$$

as a function of β as defined in Equation 4.1, where N is the number of events in the data set and π_i is the predicted probability of the i-th observed outcome:

$$\pi_i = \begin{cases} P(y_i \leq q_j|\mathbf{x}_i) & y_i \leq q_j \\ 1 - P(y_i \leq q_j|\mathbf{x}_i) & y_i > q_j \end{cases} \tag{4.3}$$

Often separate logistic regressions (i.e., with separate coefficient vectors β) are fitted for several thresholds q_j of interest (e.g., Hamill et al. 2004; Wilks 2006a; Wilks and Hamill 2007). This implies that the regression lines for different thresholds can cross, so that for some values of the predictor variables $\mathbf{x}$, $P(y \leq q_a|\mathbf{x}) > P(y \leq q_b|\mathbf{x})$ although $q_a < q_b$ which leads to nonsense negative probability for y to fall between q_a and q_b.

4.2.2 Heteroscedastic extended logistic regression (HXLR)

To avoid these negative probabilities and to reduce the number of regression coefficients Wilks (2009) proposed to include a transformation of the predictand thresholds as an additional predictor variable in logistic regression.

$$P(y \leq q_j|\mathbf{x}) = \Lambda(\alpha g(q_j) - \mathbf{x}^\top \beta) \tag{4.4}$$

where α is an additional coefficient that has to be estimated and the transformation $g()$ is a non-decreasing function. Equation 4.4 also differs from standard logistic regression, where β is estimated separately for each threshold, in that here β is the same for all thresholds. Thus, one interpretation of Equation 4.4 is that it defines parallel regression lines (in log-odds space) with equal slope but different intercepts ($\theta_j = \alpha g(q_j) - \beta_0$). Figure 4.1 shows examples of these regression lines schematically.

Extended logistic regression not only avoids the problem of crossing regression lines but also allows for computing probabilities for any threshold value q_j (and not only the the thresholds employed for estimating the model). In other words, Equation 4.4 can also be interpreted as a cumulative distribution function that describes a full continuous predictive distribution. After some reformulation (see Messner et al. 2013c), Equation 4.4 can also be written as

$$P(g(y) \leq g(q_j)|\mathbf{x}) = P(y \leq q_j|\mathbf{x}) = \Lambda\left(\frac{g(q_j) - \mathbf{x}^\top \beta/\alpha}{1/\alpha}\right) \tag{4.5}$$

which shows that the predictive distribution of the transformed predictand $g(y)$ is a logistic distribution with location parameter $\mathbf{x}^\top \beta/\alpha$ and scale parameter $1/\alpha$. Thus, the transformation $g()$ must be chosen such that the transformed predictand can be assumed to follow a conditional (on the predictors $\mathbf{x}$) logistic distribution.

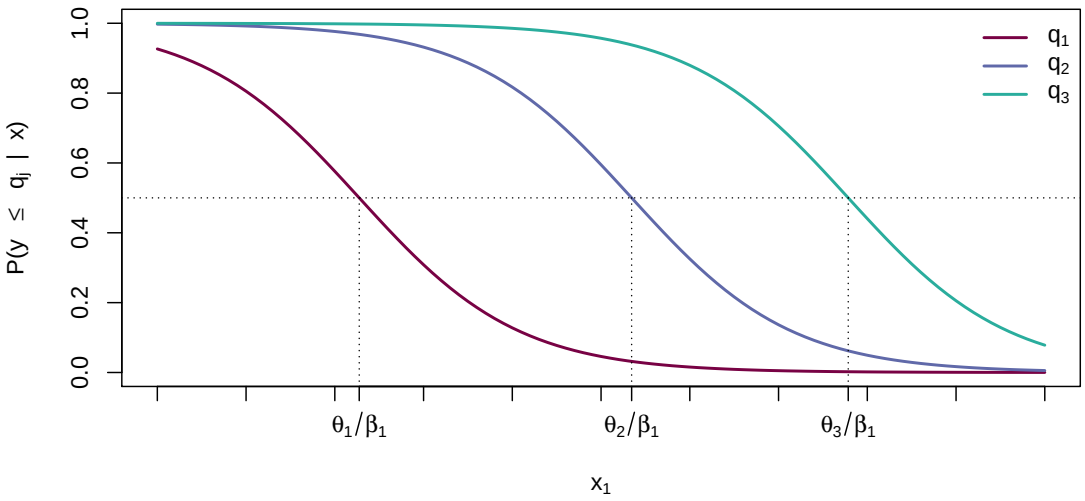

Figure 4.1: Schematic figure of regression lines fitted with extended, ordered, or censored logistic regression with one predictor variable x_1 and $J = 3$ thresholds (q_1, q_2, q_3).

To effectively utilize uncertainty information contained in the ensemble spread, Messner et al. (2013c) proposed to use additional predictor variables ($\mathbf{z} = (1, z_1, z_2, \dots)^\top$; e.g., the ensemble spread) to directly control the dispersion (variance) of the logistic predictive distribution:

$$P(y \leq q_j|\mathbf{x}) = \Lambda\left(\frac{g(q_j) - \mathbf{x}^\top\gamma}{\exp(\mathbf{z}^\top\delta)}\right) \tag{4.6}$$

where $\gamma = (1, \gamma_1, \gamma_2, \dots)^\top$ and $\delta = (1, \delta_1, \delta_2, \dots)^\top$ are the coefficient vectors that have to be estimated.

These coefficient vectors γ and δ are also estimated by maximizing the log-likelihood function given by Equation 4.2. However, the probability of the observed outcome for the multi-categorical predictand is

$$\pi_i = \begin{cases} P(y_i \leq q_1|\mathbf{x}_i) & y_i \leq q_1 \\ P(y_i \leq q_j|\mathbf{x}_i) - P(y_i \leq q_{j-1}|\mathbf{x}_i) & q_{j-1} < y_i \leq q_j \\ 1 - P(y_i \leq q_J|\mathbf{x}_i) & y_i > q_J \end{cases} \tag{4.7}$$

(Messner et al. 2013c) where J is the number of thresholds q_j that have been selected for the fitting calculation.

4.2.3 Heteroscedastic ordered logistic regression (HOLR)

Ordered logistic regression – also known as ordered logit, proportional odds logistic regression, or cumulative link model – is a popular regression model from statistics and econometrics for ordinal data, which has not received much attention in meteorology so far. Like extended logistic regression it is an extension of standard logistic regression for multi-categorical and ordered predictands. Different to extended logistic regression, separate intercepts θ_j are fitted for each selected threshold instead of modeling them as a linear function of the (transformed) thresholds.

$$P(y \leq q_j|\mathbf{x}) = \Lambda(\theta_j - \mathbf{x}^\top \beta) \tag{4.8}$$

where the estimated separate intercepts θ_j are only constrained to be ordered ($\theta_1 \leq \theta_2 \leq \cdots \leq \theta_J$). Because the intercepts of the regression lines are fully determined by θ_j further intercepts are not needed anymore so that $\mathbf{x} = (x_1, x_2, \dots)^\top$ must not contain any constant.

The separate intercepts for each threshold imply the estimation of more coefficients than for extended logistic regression. Furthermore only the probabilities for the thresholds q_j employed in the estimation can be derived, so that Equation 4.8 does not specify full continuous predictive distributions. In return, ordered logistic regression does not assume a continuous distribution for the transformed predictand. Thus no (possibly non-existent) transformation has to be determined to fulfill this assumption.

Similar to heteroscedastic extended logistic regression, a heteroscedastic version of ordered logistic regression also allows control of the scale (variance) of an underlying latent distribution with additional predictor variables (Agresti 2002).

$$P(y \leq q_j|\mathbf{x}) = \Lambda\left(\frac{\theta_j - \mathbf{x}^\top \beta}{\exp(\mathbf{z}^\top \delta)}\right) \tag{4.9}$$

Note that here also no constant is needed in $\mathbf{z} = (z_1, z_2, \dots)^\top$.

Maximum likelihood estimation with the same log-likelihood function as for extended logistic regression (Equations 4.2 and 4.7) is used to estimate the coefficients θ_j, β, and δ.

4.2.4 Heteroscedastic censored logistic regression (HCLR)

Above we have shown that extended logistic regression assumes a conditional logistic distribution for the transformed predictand. The maximum likelihood estimation with the log-likelihood function given by Equations 4.2 and 4.7 fits the selected category probabilities. However, if the predictand is given in continuous

form, the model described by Equation 4.6 can also be estimated with the log-likelihood function from Equation 4.2 with

$$\pi_i = \lambda\left(\frac{g(y_i) - \mathbf{x}^\top\gamma}{\exp(\mathbf{z}^\top\delta)}\right) = \frac{\exp\left(-\frac{g(y_i)-\mathbf{x}^\top\gamma}{\exp(\mathbf{z}^\top\delta)}\right)}{\exp(\mathbf{z}^\top\delta)\left[1+\exp\left(-\frac{g(y_i)-\mathbf{x}^\top\gamma}{\exp(\mathbf{z}^\top\delta)}\right)\right]^2} \tag{4.10}$$

where $\lambda(\cdot)$ denotes the likelihood function of the standard logistic distribution. The likelihood is notationally identical to the probability density function (i.e., the derivative of Equation 4.6 with respect to $g(q_j)$), but differs because it is a function of the parameter vectors γ and δ for a fixed predictand value y_i, rather than being a function of y_i given fixed values for γ and δ. In this way, the π_i employed for fitting the model are not the likelihoods for predictands falling into discrete intervals, but rather the likelihoods that they take on their exact observed values. This model can also be interpreted as a linear regression model with a (heteroscedastic) logistic error distribution.

Non-negative variables, e.g., wind speeds or precipitation amounts, are only continuous for positive values and have a natural threshold at 0. This non-negativity can easily accommodated using censored regression (first discussed by Tobin 1958, for the Gaussian case) where the π_i are replaced by

$$\pi_i = \begin{cases} \Lambda\left(\frac{g(0)-\mathbf{x}^\top\gamma}{\exp(\mathbf{z}^\top\delta)}\right) & y_i = 0 \\ \lambda\left(\frac{g(y_i)-\mathbf{x}^\top\gamma}{\exp(\mathbf{z}^\top\delta)}\right) & y_i > 0 \end{cases} \tag{4.11}$$

in Equation 4.2.

This heteroscedastic censored logistic regression fits a logistic error distribution with point mass at zero to the transformed predictand. While such an error distribution seems reasonable for square root transformed precipitation amounts (Scheuerer 2013; Schefzik et al. 2013), usually other error distributions are assumed for wind speed. For example Thorarinsdottir and Gneiting (2010) proposed to fit a truncated normal distribution to the *untransformed* wind speed. In this case, in Equations 4.6 and 4.10 the logistic distribution is replaced with a truncated normal distribution and $g(y)$ is set to $g(y) = y$. Note that Thorarinsdottir and Gneiting (2010) called this model also heteroscedastic censored regression although actually the data is considered to be truncated and not censored. In the following we therefore denote this model as heteroscedastic truncated Gaussian regression (*HTGR*) which we also employ as benchmark model for wind speed.

4.2.5 Comparison

Table 4.1 summarizes the major differences between the 4 different logistic regression models that were presented above. Extended logistic regression (*XLR*)

Model	*SLR*	*(H)OLR*	*(H)XLR*	*(H)CLR*
Type	Separate	Ordered	Extended	Censored
Intercepts	Unconstrained	Ordered	Lin. fun. of $g(q)$	Lin. fun. of $g(q)$
Slopes	Separate	Joint	Joint	Joint
Number of parameters	KJ	$K + J$	$K + 2$	$K + 2$
Implies cont. distribution	No	No	Yes	Yes
Estimation based on	Thresholds	Thresholds	Thresholds	Cont. distribution
Heteroscedasticity	No	Yes (optionally)	Yes (optionally)	Yes (optionally)

Table 4.1: Overview over the different logistic regression models with respect to their parametrization and the likelihood. K is the number of predictor variables $(x_1, x_2, \ldots, z_1, z_2, \ldots)$ and J is the number of thresholds q_j.

and censored logistic regression (*CLR*) (and their heteroscedastic versions HXLR and HCLR, respectively) are essentially the same models and only differ in their parameter estimation. They have the fewest parameters of the compared models but imply continuous distribution assumptions. Ordered logistic regression (*OLR*) and its heteroscedastic version (HOLR) avoid this continuous distribution assumption but require estimation of more coefficients than *(H)XLR* and *(H)CLR*. With its unconstrained slope estimates, separate logistic regressions *SLR* is more flexible than *OLR* but requires estimation of even more coefficients. Figure 4.1 shows schematic parallel regression lines for *XLR*, *CLR*, or *OLR*. In contrast to these models, regression lines from *SLR* would not be constrained to be parallel and so could potentially cross, which would lead to nonsense negative probabilities.

4.3 Data

To compare the presented ensemble post-processing methods, we used 10 meter wind speed observations (10 minute average) and 24-h accumulated precipitation amount from the two European weather stations *Paris–Orly* (48.717 N, 2.383 E) and *Wien–Hohe-Warte* (48.249 N, 16.356 E). As input for the statistical models, 10m wind speed and total precipitation ensemble forecasts from the European Centre for Medium Range Weather Forecasts (ECMWF) were linearly interpolated from neighboring grid points to the station locations. The data were available from April 2010 to December 2012 and we mainly considered the lead times 24, 48, and 96 hours for this study.

Since the predictands were square root transformed for most regression models (see Section 4.4) we mainly used the mean and standard deviation of *square root transformed* ensemble forecasts as predictor variables. For *HTGR* the untransformed predictand is used, following Thorarinsdottir and Gneiting (2010). Consequently we employed the mean and standard deviation of the *untransformed*

Model		$g(y)$	$\mathbf{x}$	$\mathbf{z}$
SLR	Separate logistic regressions	–	$(1, M, S * M)^\top$	–
HOLR	Het. ordered logistic regression	–	M	S
HXLR	Het. extended logistic regression	$\sqrt{y}$	$(1, M)^\top$	$(1, S)^\top$
HCLR	Het. censored logistic regression	$\sqrt{y}$	$(1, M)^\top$	$(1, S)^\top$
HTGR	Het. truncated Gaussian regression	y	$(1, M_r)^\top$	$(1, S_r)^\top$

Table 4.2: List of different statistical models. $g(y)$ is the transformation, $\mathbf{x}$ are vectors of predictor variables for the location (mean) and $\mathbf{z}$ are predictor variables for the scale (variance). M and S are the mean and standard deviation of square root transformed ensemble forecasts respectively and M_r and S_r are the mean and standard deviation of the untransformed ensemble forecasts respectively. For wind speed forecasts M, S, M_r, and S_r are derived from 10m wind speed ensemble forecasts and for precipitation forecasts M and S are derived from total precipitation ensemble forecasts.

ensemble forecasts as input for this model.

As thresholds q_j we defined $M = 9$ climatological deciles that are estimated for each location and predictand variable separately. Note that for precipitation several deciles are 0 and are merged to one threshold.

We found the ensemble spread to improve the forecasts of all statistical models, indicating useful spread-skill relationships. Therefore we only show results for the heteroscedastic models in the following. For separate logistic regressions the product of ensemble mean and spread is included as additional predictor variable (Wilks and Hamill 2007). Table 4.2 lists the different models that are compared in the following in detail.

4.4 Results

Before comparing the performance of the different ensemble post-processing methods we show how ordered logistic regression can be used to determine appropriate transformations $g()$ for extended logistic regression. The crosses and plus-signs in Figure 4.2 show the fitted intercepts from ordered logistic regression (*HOLR*) for the two locations and two predictands. For both locations and variables these plots suggest that the intercepts can be parameterized as being proportional to the square roots of the thresholds. Thus we fitted *HXLR* models with $g(q_j) = \sqrt{q_j}$ and added the corresponding *HXLR* intercept functions $\theta_j = \beta_0 + \alpha\sqrt{q_j}$ as lines in Figure 4.2. For both predictand variables and locations the *HXLR* intercept functions fit the *HOLR* intercepts reasonably well. Note

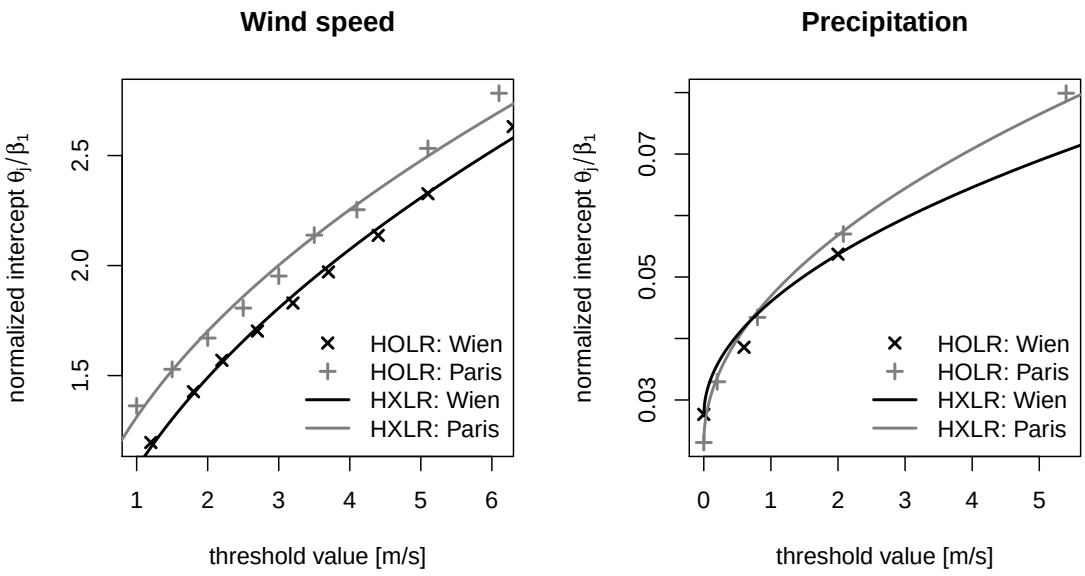

Figure 4.2: Intercepts θ_j from heteroscedastic ordered (*HOLR*) and extended logistic regression (*HXLR*) relative to threshold values, for the locations *Wien–Hohe-Warte* and *Paris–Orly*, lead time 48 hours, and the predictands wind speed (left) and 24 hours accumulated precipitation amount (right). For better comparability intercepts are normalized with β_1 respectively. The square root is used as transformation for *HXLR* ($g(q_j) = \sqrt{q_j}$).

that similar figures can also be used to compare the intercepts of extended logistic regression with those of separate logistic regression (e.g., Ruiz and Saulo 2012). However, the varying slope coefficients then complicate the comparison.

Figure 4.2 already suggests that *HXLR* and *HOLR* predict similarly well. In the following we compare these and the other statistical models more thoroughly. Because all models provide probabilistic forecasts for discrete intervals we mainly employ the ranked probability score (RPS; Epstein 1969; Wilks 2006b) to characterize forecast accuracy:

$$RPS = \sum_{j=1}^{J} (P(y \leq q_j|\mathbf{x}) - I(y \leq q_j))^2 \tag{4.12}$$

where J is the number of thresholds and $I(\cdot)$ is the indicator function. For each model, forecast location, and lead time we applied 10-fold cross validation to get independent training and test data sets. To estimate the sampling distribution for the average $\overline{RPS}$ we computed means of 250 bootstrap samples. To compare the models with a reference model we finally computed ranked probability *skill*

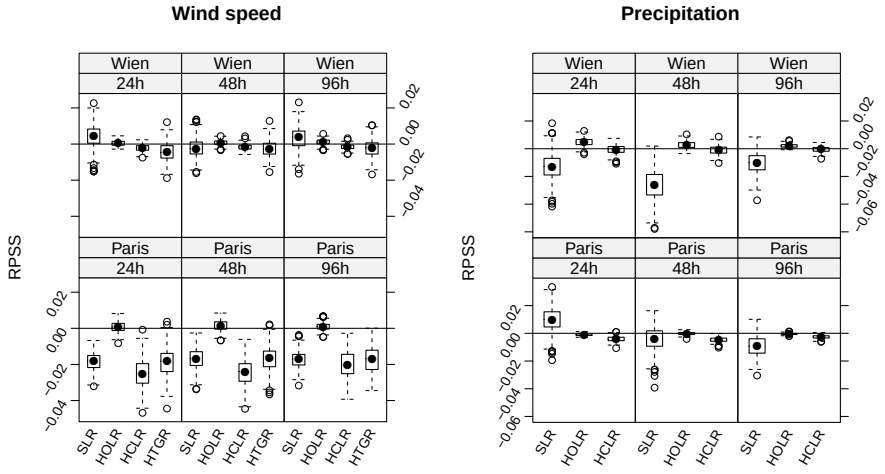

Figure 4.3: Ranked probability skill score (*RPSS*) relative to heteroscedastic extended logistic regression (*HXLR*) of wind speed (left) and 24 hours accumulated precipitation amount (right) for different models (see Table 4.2 for details) and locations. 9 climatological deciles that were computed separately for each forecast location are used as thresholds. Positive values indicate improvements over *HXLR*. The solid circles mark the median and the boxes the interquartile ranges of the 250 values from the bootstrapping approach, the whiskers show the most extreme values that are less than 1.5 times the length of the box away from the box, and empty circles are plotted for values that are outside the whiskers.

scores (*RPSS*):

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}_{ref}} \tag{4.13}$$

where $\overline{RPS}_{ref}$ is the $\overline{RPS}$ of appropriate reference forecasts.

Figure 4.3 shows the *RPSS* relative to *HXLR* for different models, lead times, locations, and predictand variables. *HOLR* performs equally well or slightly better than *HXLR* for all locations, lead times, and predictand variables. For precipitation in *Paris* forecasts of *HXLR* and *HOLR* are nearly identical which is consistent with Figure 4.2 where the *HXLR* intercept function almost perfectly interpolates the *HOLR* intercepts. Separate logistic regressions (*SLR*) mostly performs worse than *HXLR*. Exceptions are wind speed forecasts in *Wien* for 24 and 96 hours lead time and precipitation forecasts in *Paris* for 24 hours lead time. However, note that the *RPS* (Equation 4.12) does not penalize the partly inconsistent forecasts from *SLR*. *HCLR* and *HTGR* also tend to perform worse than *HXLR*. While for *Paris HTGR* is slightly better than *HCLR* there is no clear preference for one of these models in *Wien*.

Because the different statistical models differ considerably in their number

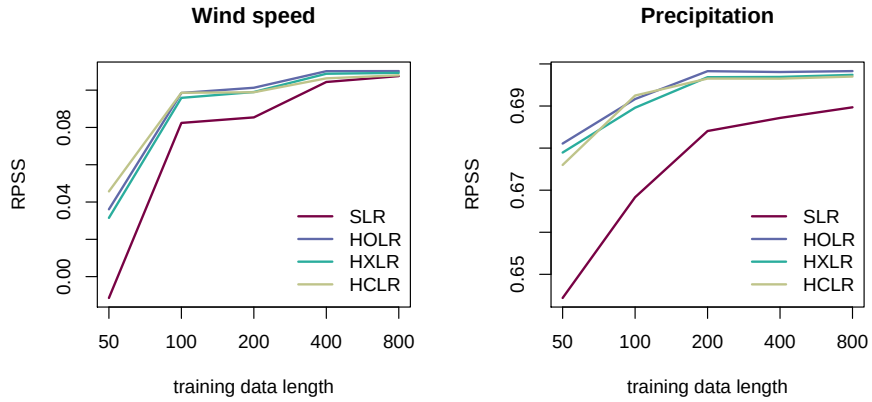

Figure 4.4: Ranked probability skill score (*RPSS*) relative to the raw ensemble (ensemble relative frequencies within each interval) in *Wien–Hohe-Warte* for different training data lengths and models (see Table 4.2 for details) and lead time 48 hours. 9 climatological deciles are used as thresholds for wind speed and 3 for precipitation.

of estimated coefficients (*SLR*: $3J$, *HOLR*: $2 + J$, *HXLR*, *HCLR*, *HTGR*: 4) it is also interesting to compare their performance for different training data lengths. Figure 4.4 shows RPSS for windspeed and precipitation forecasts for 48 hours lead time at Wien, relative to the raw ensemble interval relative frequencies. It can be seen that all models lose skill with a reduced training data set. With the largest parameter count *SLR* clearly loses most and for wind speed even performs worse than the raw ensemble ($RPSS < 0$) when the training data contains only 50 days. The other models exhibit comparable skill reductions in response to decreasing training data.

HCLR basically fits the same model as *HXLR*, with the only difference being that the estimated model parameters optimize either the selected category probabilities (*HXLR*) or the continuous predictive distribution (*HCLR*). Since the *RPS* only measures the quality of the selected category probabilities the better *RPS* of *HXLR* in Figure 4.3 is not surprising. To compare also the quality of the full predictive distributions we therefore employ the continuous ranked probability score (*CRPS*; Matheson and Winkler 1976; Hersbach 2000; Wilks 2006b) that generalizes the *RPS* to full predictive distributions.

$$CRPS = \int_{-\infty}^{\infty} (P(y_i \leq t|\mathbf{x}) - I(y_i \leq t))^2 dt \tag{4.14}$$

Analogously to Figure 4.3 the continuous ranked probability skill score (*CRPSS*) relative to *HXLR* is shown in Figure 4.5. In contrast to the *RPSS* (Figure 4.3) the *CRPSS* clearly favors *HCLR* for both locations and predictand variables.

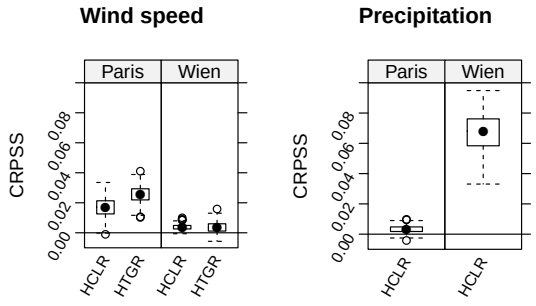

Figure 4.5: Continuous ranked probability skill score (*CRPSS*) relative to heteroscedastic extended logistic regression (*HXLR*) and their bootstrap sampling distributions in *Wien–Hohe-Warte* for different predictands, models (see Table 4.2 for details), and locations. Positive values indicate improvements over *HXLR*.

Note that the large improvement of *HCLR* over *HXLR* for precipitation in *Wien* mainly stems from *HXLR*'s bad forecast performance for very high precipitation amounts. The inclusion of additional thresholds in the parameter fitting process (e.g., climatological 0.95-quantile) substantially improved the *CRPS* of *HXLR* and consequently diminished the *CRPSS* of *HCLR* (not shown).

For wind speed, Figure 4.5 also shows the *CRPSS* for *HTGR*. As in Figure 4.3 *HCLR* and *HTGR* show similar *CRPSS* for *Wien* while *HTGR* is slightly preferred for *Paris*, which suggests that there the real error distribution is better estimated by a truncated normal than by a censored transformed logistic distribution.

Finally Figures 4.6 and 4.7 show reliability diagrams (e.g., Wilks 2006b) for the lower and upper climatological deciles, respectively, for 48 hours lead time at *Wien*. With few exceptions the observed conditional relative frequencies of both predictand variables lie within the 95% consistency intervals (Bröcker and Smith 2007) with only minor differences between the different statistical models. Similarly, the refinement distributions in Figures 4.6 and 4.7 show only little differences between the different models. Only for zero precipitation *SLR* and *HOLR* have slightly sharper forecasts than *HXLR* and *HCLR* (forecasts more frequently close to 0 and 1).

4.5 Summary and conclusion

Extended logistic regression fits predictand category probabilities by assuming a conditional logistic distribution for the transformed predictand (Scheuerer 2013; Schefzik et al. 2013; Messner et al. 2013c). However, for some applications the transformed predictand cannot be assumed to follow a logistic distribution.

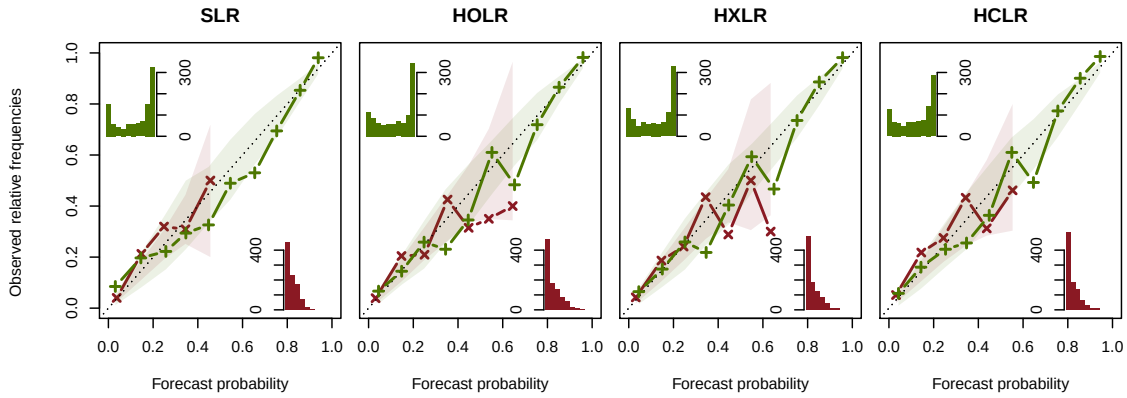

Figure 4.6: Reliability diagrams for predicted probabilities to fall below the first climatological decile $P(y \leq q_1|\mathbf{x})$ for *Wien–Hohe-Warte*, lead time 48 hours, and different models. Forecasts are aggregated in 0.1 probability intervals. Calibration functions for *wind speed* are plotted as red 'x' and for *precipitation amount* as green '+' and are only shown for intervals with more than 10 forecasts. Refinement distributions for wind speed are plotted in the bottom right corner in red and for precipitation in the top left corner in green. 95% consistency intervals derived from consistency re-sampling (Bröcker and Smith 2007) are shown as red and green shaded areas respectively. Note that due to the frequent zero observations $q_1 = q_2 = \cdots = q_6$ for precipitation so that $P(y \leq q_1|\mathbf{x}) = P(y \leq q_6|\mathbf{x}) = P(y = 0|\mathbf{x})$.

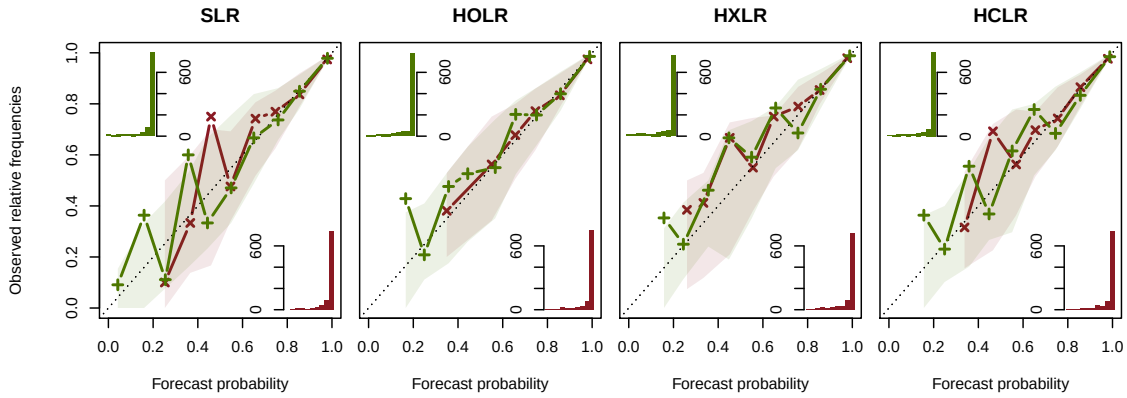

Figure 4.7: Same as Figure 4.6 but for predicted probabilities to fall below the upper climatological decile $P(y \leq q_9|\mathbf{x})$.

Moreover, fitting selected category probabilities implies disregarding available information when the predictand is actually given in continuous form.

In this study we compared extended logistic regression with two closely related regression models from statistics and econometrics. Ordered logistic regression is very similar to extended logistic regression but avoids a continuous distribution assumption. On the other hand, censored logistic regression fits the same model as extended logistic regression but uses each individual predictand value in the training data set instead of the selected category probabilities. As further benchmark models we also employed separate logistic regressions and a truncated Gaussian regression model (Thorarinsdottir and Gneiting 2010). The performance of the different statistical models was tested with wind speed and precipitation data from two European locations and ensemble forecasts from the ECMWF. Overall, the logistic distribution assumption seemed to be quite appropriate for the square-root-transformed predictands, at both locations and for both predictand variables. Thus, the performance differences between ordered and extended logistic regression were only minor. However, because no continuous distribution has to be assumed, ordered logistic regression should generally be preferred if solely threshold probabilities are required.

Since extended logistic regression fits selected category probabilities, it is actually not surprising that *RPS* skills are higher for this model than for censored logistic regression, which fits the full continuous predictive distribution. For the same reason it is unsurprising that censored logistic regression performed better than extended logistic regression according to *CRPS* skill, which evaluates accuracy of the full predictive distributions.

Extended and censored logistic regression assume censored conditional logistic distributions for the transformed predictand. In contrast, wind speed was assumed to follow a truncated normal distribution in Thorarinsdottir and Gneiting (2010). A comparison between censored and truncated regression models showed that the assumption of a truncated normal distribution resulted in slightly better wind speed forecasts than the assumption of a censored transformed logistic distribution.

As input for the statistical methods we only employed NWP model forecasts for the observation location and forecast variable. However, all models might potentially be improved with additional inputs such as NWP model forecasts of other variables and/or locations.

Nevertheless, our results show that the optimal statistical model strongly depends on the intended application. Ordered logistic regression was best suited for category probability predictions for the forecasts considered here, given sufficiently long training series. When the transformed predictand can be assumed

to follow a conditional logistic distribution then extended logistic regression provides equally good category probability forecasts while requiring fewer coefficients and additionally specifying full predictive distributions. However, if the primary interest is in predicting full continuous probability distributions, censored or truncated regression models should be preferred because they use the information contained in the training data more fully.

Acknowledgements

This study was supported by the Austrian Science Fund (FWF): L615-N10. The first author was also supported by a PhD scholarship from the University of Innsbruck, *Vizerektorat für Forschung*. Data from the ECMWF forecasting system were obtained from the ECMWF Data Server.

A: Computational details

Our results were obtained on Ubuntu Linux using the statistical software R 2.15.2 (R Core Team 2013). Heteroscedastic extended logistic regression and heteroscedastic censored logistic regression were fitted using the package crch 0.1-0 (Messner and Zeileis 2013). For ordered logistic regression models we used the package ordinal 2012.09-11 (Christensen 2013).

5

Summary and conclusions

The highly volatile nature of wind complicates the grid-integration of the steadily increasing amount of wind power. To cope with this volatility grid operators and managers strongly demand accurate wind and wind power forecasts. Probabilistic forecasts that allow them to estimate the forecast uncertainty are often most valuable. Many weather forecasting centers nowadays compute ensemble forecasts to provide forecast uncertainty information. However, these ensemble forecasts are often uncalibrated and not tailored to wind power forecasts. In this thesis we tested, developed, and improved methods to statistically post-process ensemble forecasts for probabilistic wind and wind power predictions.

In the first part of this thesis, we proposed a novel approach to deal with the conversion problem from wind to wind power. In this approach the inverse of the power-curve function is used to transform wind *power* into wind *speed*. The limited range of power production between zero and nominal power can then easily be respected with censored regression models. With this combined strategy, NWP output can be post-processed with simple linear parametric regression models. A comparison with more complex nonlinear and nonparametric regression models showed that although much simpler, our approach performed similarly well and is especially attractive for small training data sets.

Extended logistic regression (Wilks 2009) is a popular ensemble post-processing method that fits a conditional *censored* logistic distribution to the (transformed) predictand (Scheuerer 2013; Schefzik et al. 2013; Messner et al. 2013c). Thus, it is a censored regression model and therefore well-suited to be used in combination with the inverse power-curve transformation. Traditionally,

extended logistic regression has mainly been used for precipitation forecasts. In the second and third part of this thesis we improved extended logistic regression and tested its suitability for wind speed forecasts.

Although extended logistic regression has mainly been used for ensemble post-processing, ensemble spread information was often neglected because it did not improve the forecasts. We found that in the original formulation of extended logistic regression, ensemble spread information can only be used to control the location (mean) but not the dispersion (variance) of the logistic predictive distribution. Uncertainty information contained in the ensemble spread is therefore not utilized appropriately. To overcome this problem we proposed a heteroscedastic extended logistic regression approach where the ensemble spread is directly used as predictor for the dispersion of the logistic predictive distribution. A case study with wind speed data from several European weather stations and ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) showed that this new approach utilizes ensemble spread information effectively to improve probabilistic wind speed forecasts from extended logistic regression.

Several recent studies noticed that extended logistic regression assumes a conditional (censored) logistic distribution for the (transformed) predictand (Scheuerer 2013; Schefzik et al. 2013; Messner et al. 2013c) where this distribution is fitted to selected predictand category probabilities. However, a logistic distribution assumption might not be appropriate for all applications. Furthermore, fitting selected *category* probabilities implies disregarding available information of *continuous* predictands. In the third part of this thesis we compared extended logistic regression with two closely related popular regression models from statistics and econometrics that have not received much attention in meteorology so far. Ordered logistic regression also fits selected predictand category probabilities but avoids a continuous distribution assumption. Censored logistic regression, similar to the model that was proposed in the first part of this thesis, fits the same model as extended logistic regression but uses each individual predictand value in the training data set instead of selected category probabilities. With wind speed and precipitation data from two European locations, we showed that the suitability of these methods strongly depends on the intended application. Category probabilities are best predicted by ordered logistic regression whereas full probability forecasts are better estimated by censored logistic regression. Extended logistic regression is a compromise of both approaches with reasonably well predicted category probabilities and full predictive distributions. Furthermore, a comparison with heteroscedastic truncated Gaussian regression (Thorarinsdottir and Gneiting 2010) suggested that the conditional censored lo-

gistic distribution assumption of extended and censored logistic regression is appropriate for square rooted wind speed.

For simplicity, we only employed NWP wind speed ensemble forecasts, interpolated to the forecast location, as input for the different wind speed and wind power forecasting models. However, in future studies and especially in an operational setting the forecasts might be improved by adding further input variables like current observations, other NWP forecast variables or NWP forecasts from other locations and/or model levels.

In addition to wind speed data we also investigated precipitation data in the third part of this thesis. This already indicates that the findings of this thesis are not only relevant for wind and wind power predictions. Especially the results of the last two parts can easily be transferred to other meteorological variables.

The three parts of this thesis are all treated in separate scientific articles. Two are already in press (Messner et al. 2013a,c) and one is submitted (Messner et al. 2013b) to international peer-reviewed journals. This guarantees our findings to find their way into the wind power and ensemble post-processing communities. Furthermore, we released an R software package (Messner and Zeileis 2013; R Core Team 2013) with functions to fit our proposed statistical models. This facilitates testing and improving these models in future studies or operational implementations.

Bibliography

Agresti, A., 2002: *Categorical Data Analysis*. 2d ed., John Wiley & Sons, 734 pp.

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting*, **28 (2)**, 515–524.

Bessa, R. J., V. Miranda, and A. Botterud, 2012a: Time adaptive conditional kernel density estimation for wind power forecasting. *IEEE Transactions on Sustainable Energy*, **3 (4)**, 660–669.

Bessa, R. J., V. Miranda, A. Botterud, Z. Zhou, and J. Wang, 2012b: Time-adaptive quantile-copula for wind power probabilistic forecasting. *Renewable Energy*, **40 (1)**, 29–39.

Betz, A., 1920: Das maximum der theoretisch mglichen ausnutzung des windes durch windmotoren. *Zeitschrift fr das gesamte Turbinenwesen*, **26**, 307–309.

Bremnes, J. B., 2004: Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, **7 (1)**, 47–54.

Bremnes, J. B., 2006: A comparison of a few statistical models for making quantile wind power forecasts. *Wind Energy*, **9 (1-2)**, 3–11.

Bröcker, J. and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Weather and Forecasting*, **22 (3)**, 651–661.

Cabezon, D., I. Marti, M. J. San-Isidro, and I. Perez, 2004: Comparison of methods for power curve modeling. *Proceedings to Global Windpower 2004*, Chicago.

Christensen, R. H. B., 2013: *ordinal: Regression Models for Ordinal Data*. URL `http://CRAN.R-project.org/package=ordinal`, R package version 2013.09-30.

Drechsel, S., G. J. Mayr, J. W. Messner, and R. Stauffer, 2012: Wind speeds at heights crucial for wind energy: Measurements and verification of forecasts. *Journal of Applied Meteorology and Climatology*, **51 (9)**, 1602–1617.

Efron, B. and R. J. Tibshirani, 1994: *An Introduction to the Bootstrap*. Chapman and Hall.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, **8 (6)**, 985–987.

Giebel, G., J. Badger, and L. Landberg, 2005: Wind power prediction using ensembles. Tech. rep., Risø National Laboratory, 43 pp.

Giebel, G., R. Brownsword, G. N. Kariniotakis, M. Denhard, and C. Draxl, 2011: The state-of-the-art in short-term prediction of wind power – a literature overview. Tech. rep., ANEMOS. plus, 109 pp.

Glahn, H. and D. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, **11 (8)**, 1203–1211.

Gneiting, T. and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102 (477)**, 359–378.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118.

Hamill, T. M., 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous united states. *Monthly Weather Review*, **140 (7)**, 2232–2252.

Hamill, T. M., C. Snyder, and J. S. Whitaker, 2003: Ensemble forecasts and the properties of flow-dependent analysis-error covariance singular vectors. *Monthly Weather Review*, **131 (8)**, 1741–1758.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, **132 (6)**, 1434–1447.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15 (5)**, 559–570.

Hothorn, T., F. Leisch, A. Zeileis, and K. Hornik, 2005: The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, **14 (3)**, 675–699.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1995: *Continuous Univariate Distributions*, Vol. 2. John Wiley & Sons, 752 pp.

Juban, J., L. Fugon, and G. N. Kariniotakis, 2007: Probabilistic short-term wind power forecasting based on kernel density estimators. *Proceedings of the 2007 European Wind Energy Conference*, Milan, 7–10.

Koenker, R., 2012: *quantreg: Quantile Regression*. URL `http://CRAN.R-project.org/package=quantreg`, R package version 4.91.

Koenker, R. and G. Bassett Jr, 1978: Regression quantiles. *Econometrica*, **46 (1)**, 33–50.

Landberg, L., 1999: Short-term prediction of the power production from wind farms. *Journal of Wind Engineering and Industrial Aerodynamics*, **80 (1-2)**, 207–220.

Lange, M., 2005: On the uncertainty of wind power predictions – analysis of the forecast accuracy and statistical distribution of errors. *Journal of Solar Energy Engineering*, **127 (2)**, 177–184.

Lin, G., X. He, and S. Portnoy, 2012: Quantile regression with doubly censored data. *Computational Statistics & Data Analysis*, **56 (4)**, 797–812.

Lorenz, E., 1996: Predictability: A problem partly solved. *Proceedings of the ECMWF Seminar on Predictability*, 1–18.

Louka, P., G. Galanis, N. Siebert, G. N. Kariniotakis, P. Katsafados, I. Pytharoulis, and G. Kallos, 2008: Improvements in wind speed forecasts for wind power prediction purposes using kalman filtering. *Journal of Wind Engineering and Industrial Aerodynamics*, **96 (12)**, 2348–2362.

Matheson, J. E. and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management Science*, **22 (10)**, 1087–1096.

Messner, J. W. and A. Zeileis, 2013: *crch: Censored Regression with Conditional Heteroscedasticity*. URL `http://CRAN.R-project.org/package=crch`, R package version 0.1-0.

Messner, J. W., A. Zeileis, J. Broecker, and G. J. Mayr, 2013a: Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. *Wind Energy*, in press.

Messner, J. W., A. Zeileis, G. J. Mayr, and D. S. Wilks, 2013b: Extending extended logistic regression for ensemble post-processing: Extended vs. separate vs. ordered vs. censored. *Monthly weather review*, submitted.

Messner, J. W., A. Zeileis, G. J. Mayr, and D. S. Wilks, 2013c: Heteroscedastic extended logistic regression for post-processing of ensemble guidance. *Monthly Weather Review*, in press.

Møller, J. K., H. A. Nielsen, and H. Madsen, 2008: Time-adaptive quantile regression. *Computational Statistics & Data Analysis*, **52 (3)**, 1292–1303.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122 (529)**, 73–119.

Müller, M. D., 2011: Effects of model resolution and statistical postprocessing on shelter temperature and wind forecasts. *Journal of Applied Meteorology and Climatology*, **50 (8)**, 1627–1636.

Nelder, J. A. and R. W. M. Wedderburn, 1972: Generalized linear models. *Journal of the Royal Statistical Society A*, **135 (3)**, 370–384.

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 (1-2)**, 95–108.

Nielsen, H. A., H. Madsen, T. S. Nielsen, J. Badger, G. Giebel, L. Landberg, K. Sattler, and H. Feddersen, 2004: Wind power ensemble forecasting. *Proceedings of the 2004 Global Wind Power Conference*, Chicago.

Nielsen, H. A., T. S. Nielsen, H. Madsen, M. J. S. I. Pindado, and I. Marti, 2007: Optimal combination of wind power forecasts. *Wind Energy*, **10 (5)**, 471–482.

Nielsen, T. S., H. Madsen, and H. A. Nielsen, 2001: Zephyr – the prediction models. *Proceedings of the 2001 European Wind Energy Conference*, Copenhagen.

Peng, L. and Y. Huang, 2008: Survival analysis with quantile regression models. *Journal of American Statistical Association*, **103**, 637–649.

Pinson, P., 2004: Short-term wind power prediction for offshore wind farms – evaluation of fuzzy-neural network based models. *Proceedings of the 2004 Global Wind Power Conference*, Chicago.

Pinson, P., 2012: Adaptive calibration of (u,v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138 (666)**, 1273–1284.

Pinson, P. and G. Kariniotakis, 2004: On-line assessment of prediction risk for wind power production forecasts. *Wind Energy*, **7 (2)**, 119–132.

Pinson, P. and G. Kariniotakis, 2010: Conditional prediction intervals of wind power generation. *IEEE Transactions on Power Systems*, **25 (4)**, 1845–1856.

Pinson, P. and H. Madsen, 2009: Ensemble-based probabilistic forecasting at Horns Rev. *Wind Energy*, **12 (2)**, 137–155.

Pinson, P., H. A. Nielsen, J. K. Møller, H. Madsen, and G. N. Kariniotakis, 2007: Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy*, **10 (6)**, 497–516.

Portnoy, S., 2003: Censored quantile regression. *Journal of American Statistical Association*, **98**, 1001–1012.

Powell, J. L., 1986: Censored regression quantiles. *Journal of Econometrics*, **32 (1)**, 143–155.

R Core Team, 2013: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL `http://www.R-project.org/`.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133 (5)**, 1155–1174.

Roulin, E. and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review*, **140 (3)**, 874–888.

Roulston, M., 2003: Using medium-range weather forcasts to improve the value of wind energy production. *Renewable Energy*, **28 (4)**, 585–602.

Roulston, M. S., D. T. Kaplan, J. Hardenberg, and L. A. Smith, 2001: Value of the ECMWF ensemble prediction system for forecasting wind energy production. *Proceedings of the 2001 European Wind Energy Conference*, Copenhagen.

Roulston, M. S. and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus A*, **55 (1)**, 16–30.

Ruiz, J. J. and C. Saulo, 2012: How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: Sensitivity to calibration methods. *Meteorological Applications*, **19 (3)**, 302–313.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, in press.

Scheuerer, M., 2013: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, in press.

Schmeits, M. J. and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, **138 (11)**, 4199–4211.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Monthly Weather Review*, **140 (10)**, 3204–3219.

Simmons, A. and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **128 (580)**, 647–677.

Taylor, J. W., P. E. McSharry, and R. Buizza, 2009: Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, **24 (3)**, 775–782.

Thorarinsdottir, T. L. and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society A*, **173 (2)**, 371–388.

Tobin, J., 1958: Estimation of relationships for limited dependent variables. *Econometrica*, **26 (1)**, 24–36.

Wang, X. and C. H. Bishop, 2003: A comparison of breeding and ensemble transform kalman filter ensemble forecast schemes. *Journal of the Atmospheric Sciences*, **60 (9)**, 1140–1158.

Wang, X. and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, **131 (607)**, 965–986.

Wilkes, J. and J. Moccia, 2013: Wind in power – 2012 European statistics. Tech. rep., European Wind Energy Association (EWEA). URL `http://www.ewea.org/statistics/european/`.

Wilkes, J., J. Moccia, and M. Dragan, 2012: Wind in power – 2011 European statistics. Tech. rep., European Wind Energy Association (EWEA). URL `http://www.ewea.org/statistics/european/`.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13 (3)**, 243–256.

Wilks, D. S., 2006b: *Statistical Methods in the Atmospheric Sciences*, Vol. 14. 2d ed., Academic Press, 627 pp.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, **368 (March)**, 361–368.

Wilks, D. S. and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135 (6)**, 2379–2390.

Zugno, M., T. Jónsson, and P. Pinson, 2012: Trading wind energy on the basis of probabilistic forecasts both of wind generation and of market quantities. *Wind Energy*, **16 (6)**, 909–926.

Acknowledgements

Without a doubt, the writing of this thesis would not have been possible without the support of many people.

First of all, special thanks go to my parents who always supported me in all respects and are simply the best parents I could imagine.

Secondly, I want to thank my advisors Georg J. Mayr, Achim Zeileis, and Jochen Bröcker for their essential contributions to this thesis. In particular, I want to thank Georg for his great support since I started working on my master thesis. I could not imagine a better advisor. Jochen's support was particularly important in the beginning of my PhD studies and guided me to the research issues that are treated in this thesis. I really enjoyed my time in Dresden working with him. To Achim I owe a great part of my important knowledge about statistics. Furthermore I learned a lot from his precise way of working.

I also want to thank my colleagues Susanne Drechsel, Felix Schüller and especially Reto Stauffer for our fruitful discussions and collaborations.

Furthermore I thank Daniel S. Wilks for his important contributions to two of the scientific articles that are included in this thesis.

The Austrian Science Fund (FWF, L615-N10) and a PhD scholarship from the University of Innsbruck, *Vizerektorat für Forschung* supported me financially.

Many thanks also go to my two brothers, my sister, and all of my friends and colleagues for the great times I had as PhD student. Last but not least I want to thank my girlfriend for her love and support and the great time we have had together.

Curriculum Vitae

Name: Jakob Wolfgang Messner
Born: 13 July 1985 in Rum, Austria
Address: Innrain 71, 6020 Innsbruck, Austria
Email: jakob.messner@uibk.ac.at

EDUCATION AND PROFESSIONAL TRAINING:

2009–present	Doctoral study in atmospheric science at the Institute of Meteorology and Geophysics, University of Innsbruck, Austria
2008–2012	Bachelor degree course Technical Mathematics at the University of Innsbruck.
2007–2008	One semester abroad at the Complutense University of Madrid, Spain.
2004–2009	Diploma study at the University of Innsbruck. *Master of Natural Science (Magister rerum naturalium)* in Meteorology.
1995–2003	Bundesrealgymnasium Reutte. *Matura*.

EMPLOYMENT:

2009–2012	Research assistant, Institute of Meteorology and Geophysics, University of Innsbruck
2010	Guest scientist at the Max Planck Institute for Physics of Complex Systems, Dresden, Germany

TEACHING EXPERIENCES:

2011	Lecturer for 'Wetterbesprechung' (Master degree course Atmospheric Science).
2009	Tutor for 'Stochastische Prozesse' (Master degree course Atmospheric Science).
2008–2009	Tutor for 'Einführung in die Mathematik' I and II (Bachelor degree course Geo- and Atmospheric Science).

Publications

REFEREED PUBLICATIONS:

Messner, J. W. and G. J. Mayr, 2010: Probabilistic forecasts using analogs in the idealized Lorenz96 setting. *Monthly Weather Review*, **139**, 1960–1971.

Messner, J. W., A. Zeileis, J. Broecker, and G. J. Mayr, 2013: Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. *Wind Energy*, in press.

Messner, J. W., A. Zeileis, G. J. Mayr, and D. S. Wilks, 2013: Heteroscedastic extended logistic regression for post-processing of ensemble guidance. *Monthly Weather Review*, in press.

SUBMITTED PAPERS:

Messner, J. W., A. Zeileis, G. J. Mayr, and D. S. Wilks, 2013: Extending extended logistic regression for ensemble post-processing: extended vs. separate vs. ordered vs. censored. *Monthly Weather Review*, submitted.

NON-REFEREED PAPERS:

Umlauf, N., G. Mayr, and J. Messner, 2012: Why Does It Always Rain on Me? A Spatio-Temporal Analysis of Precipitation in Austria. *Austrian Journal of Statistics*, **41**, 81–92.

CONFERENCE CONTRIBUTIONS:

Messner, J. W., J. Broecker, and G. J. Mayr, 2011: Improve probabilistic wind power forecasts through previously transforming the input data. *European Meteorological Society Meeting – European Conference on Applications of Meteorology*, Berlin.

Drechsel, S., G. J. Mayr, J. W. Messner, and R. Stauffer, 2011: Observations and verification of forecasts of wind in the lower boundary layer. *European Meteorological Society Meeting – European Conference on Applications of Meteorology*, Berlin.

Umlauf, N., G. J. Mayr, J. W. Messner, and A. Zeileis, 2012: Weekend – and the Weather is Bad Again! A Spatio-Temporal Analysis of Precipitation in Austria. *European Geosciences Union General Assembly*, Wien.

Messner, J. W. and G. J. Mayr, 2012: Probabilistic Forecasts Using Analogs in the Idealized Lorenz96 Setting. *International Conference on Ensemble Methods in Geophysical Sciences*, Toulouse.

Messner, J. W., A. Zeileis, J. Bröcker, and G. J. Mayr, 2013: Improved Probabilistic Wind Power Forecasts with an Inverse Power Curve Transformation and Censored Regression. *DACH Meteorologentagung*, Innsbruck.

Messner, J. W., A. Zeileis, G. J. Mayr, and D. S. Wilks, 2013: A New Approach to Effectively Utilize the Ensemble Spread in Extended Logistic Regression. *European Meteorological Society Meeting – European Conference on Applications of Meteorology*, Reading.

SOFTWARE PACKAGES:

Messner, J. W. and A. Zeileis, 2013: *crch: Censored Regression with Conditional Heteroscedasticity*. R package version 0.1-0.
URL `http://CRAN.R-project.org/package=crch`