University of Innsbruck

Non-homogeneous boosting for predictor selection in ensemble post-processing

Jakob W. Messner, Georg J. Mayr, Achim Zeileis

Working Papers in Economics and Statistics

2016-04

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
Research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Non-Homogeneous Boosting for Predictor Selection in Ensemble Post-Processing

Jakob W. Messner
Universität Innsbruck

Georg J. Mayr
Universität Innsbruck

Achim Zeileis
Universität Innsbruck

Abstract

Non-homogeneous regression is often used to statistically post-process ensemble forecasts. Usually only ensemble forecasts of the predictand variable are used as input but other potentially useful information sources are ignored. Although it is straightforward to add further input variables, overfitting can easily deteriorate the forecast performance for increasing numbers of input variables. This paper proposes a boosting algorithm to estimate the regression coefficients while automatically selecting the most relevant input variables by restricting the coefficients of less important variables to zero. A case study with ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) shows that this approach effectively selects important input variables to clearly improve minimum and maximum temperature predictions at 5 central European stations.

Keywords: non-homogeneous regression, variable selection, boosting, statistical ensemble post-processing.

1. Introduction

Over the past decades ensemble forecasts have become an important tool for estimating the uncertainty of numerical weather prediction models. To account for initial condition and model errors, numerical models are integrated several times with slightly different initial conditions and sometimes different parameterization schemes. However, because of insufficient representation of these errors such ensembles of predictions are often biased and do not fully represent the forecast uncertainty. Therefore ensemble forecasts are often statistically post-processed to obtain unbiased and calibrated probabilistic forecasts.

Over the past years a variety of different ensemble post-processing methods have been proposed. Aside from e.g., ensemble dressing (Roulston and Smith 2003), Bayesian model averaging (Raftery, Gneiting, Balabdaoui, and Polakowski 2005), or (extended) logistic regression (Hamill, Whitaker, and Wei 2004; Wilks 2009; Messner, Zeileis, Mayr, and Wilks 2014b), non-homogeneous regression (Gneiting, Raftery, Westveld, and Goldman 2005) is particularly popular. It assumes a parametric predictive distribution and models the distribution parameters as linear functions of predictor variables such as the ensemble mean and ensemble standard deviation. In recent years it has been used for several different forecast variables (e.g., Thorarinsdottir and Gneiting 2010; Scheuerer 2014; Scheuerer and Hamill 2015) and has been extended to account for covariance structures (Pinson 2012; Schuhen, Thorarinsdottir, and Gneiting 2012; Schefzik, Thorarinsdottir, and Gneiting 2013; Feldmann, Scheuerer,

and Thorarinsdottir 2015) or to predict full spatial fields (Scheuerer and Büermann 2014; Feldmann *et al.* 2015). In most publications only the ensemble forecast of the predictand variable was used as input for the non-homogeneous regression model. However, Scheuerer (2014) and Scheuerer and Hamill (2015) showed that additional input variables can be easily incorporated and can clearly improve the forecast performance. The set of potentially useful input variables is huge and includes, among others, ensemble forecasts for other variables or locations, deterministic forecasts, current observations, transformations and interactions of all of these. Since using too many input variables can deteriorate the forecast accuracy through overfitting, the input variables should be selected carefully. Doing this by hand can be a cumbersome task that requires expert knowledge and should be done separately for each forecast variable, station and lead time.

For post-processing of deterministic predictions, stepwise regression has commonly been used to automatically select the most important input variables (e.g., Glahn and Lowry 1972; Wilson and Vallé 2002). However, to our knowledge, automatic variable selection has not yet been used for ensemble post-processing with non-homogeneous regression. In this paper we propose a boosting algorithm to automatically select the most relevant predictor variables in non-homogeneous regression. Boosting has originally been proposed for classification problems (Freund and Schapire 1997) but has also been extended and used for regression (Friedman, Hastie, and Tibshirani 2000; Bühlmann and Yu 2003; Bühlmann and Hothorn 2007; Hastie, Tibshirani, and Friedman 2013). Like other optimization algorithms boosting finds the minimum of the loss function iteratively but in each step it only updates the coefficient that improves the current fit most. Thus, if it is stopped before convergence, only the most important predictor variables have non-zero coefficients so that less relevant variables are ignored.

To investigate this novel boosting approach and to compare its performance against ordinary non-homogeneous regression we use maximum and minimum temperature forecasts at five stations in central Europe. As potential input variables we use ensemble forecasts for different weather variables from the European Centre for Medium-Range Weather Forecasts (ECMWF).

The remainder of this paper is structured as follows: The following section describes the non-homogeneous regression approach and introduces the boosting algorithm to estimate the regression coefficients. Subsequently Section 3 describes the data that is used to compute the results that are presented in Section 4. Finally, Section 5 provides a summary and conclusion.

2. Methods

This section first describes the non-homogeneous regression approach of Gneiting *et al.* (2005) and subsequently presents a boosting algorithm to automatically select the most relevant input variables.

2.1. Non-homogeneous regression

Non-homogeneous regression, sometimes also called ensemble model output statistics, was first proposed by Gneiting *et al.* (2005) for normally distributed predictands such as temperature and sea level pressure. Later publications extended this method to variables described by non-normal distributions, e.g., wind (truncated normal: Thorarinsdottir and Gneiting 2010),

or precipitation (generalized extreme value: Scheuerer 2014, censored logistic: Messner, Mayr, Wilks, and Zeileis 2014a, or censored gamma: Scheuerer and Hamill 2015). In the following, we only regard non-homogeneous *Gaussian* regression (NGR), but all concepts can easily be transferred to other distributions as well.

NGR assumes the observations y to follow a normal distribution $\mathcal{N}$ with mean μ (location) and variance σ^2 (squared scale):

$$y \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

where the location μ and the logarithm of the scale σ are expressed as

$$\text{location:} \qquad \mu = \mathbf{x}^\top \beta \tag{2}$$

$$\text{log-scale:} \qquad \log(\sigma) = \mathbf{z}^\top \gamma \tag{3}$$

with $\mathbf{x} = (1, x_1, x_2, \ldots)^\top$ and $\mathbf{z} = (1, x_1, x_2, \ldots)^\top$ being vectors of predictor variables, and $\beta = (\beta_0, \beta_1, \beta_2, \ldots)^\top$ and $\gamma = (\gamma_0, \gamma_1, \gamma_2, \ldots)^\top$ the corresponding coefficient vectors. Note that y, $\mathbf{x}$, $\mathbf{z}$, μ, and σ are event specific but indices were omitted to enhance the readability. The logarithmic link function in Equation 3 ($\log(\sigma)$) is used to assure positive values for σ. Alternatively, often also σ^2 is modeled where all coefficients in γ are restricted to be positive (e.g., Gneiting *et al.* 2005).

The coefficients β and γ are estimated by minimizing a loss function such as the negative log-likelihood or the continuous ranked probability score (CRPS). In the following, we use the negative log-likelihood, but all concepts can be easily transfered to any other differentiable loss function as well. The negative log-likelihood (L) for a single event is given by:

$$L(\mu, \sigma) = -\log\left(\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\right) \tag{4}$$

where $\phi()$ is the probability density function of the normal distribution. The full negative log-likelihood, that is used to estimate β and γ, is derived by taking the sum of $L(\mu, \sigma)$ over the training data. We perform this optimization with the Broyden-Fletcher–Goldfarb-Shanno (BFGS) algorithm as implemented in `R` (R Core Team 2015), similar to e.g., Gneiting *et al.* (2005); Thorarinsdottir and Gneiting (2010); Scheuerer (2014). For an increased efficiency of this optimization we also use analytical gradients and Hessian matrices of the log-likelihood (Messner, Mayr, and Zeileis 2016). In most studies, $\mathbf{x}$ is a vector including different ensemble member forecasts or the ensemble mean forecast while $\mathbf{z}$ usually contains the ensemble variance or standard deviation. Scheuerer (2014) and Scheuerer and Hamill (2015) also included further input variables, however, typically only ensemble forecasts of the predictand variable have been used (e.g., only ensemble predictions of temperature are included in $\mathbf{x}$ and $\mathbf{z}$ for temperature forecasts).

Clearly, many more information sources could be used as inputs, e.g., different ensemble forecast variables, current observations, deterministic forecasts, or transformations or interactions of all of these. However, adding too many variables can easily result in overfitting so that the input variables must be selected carefully. Considering the huge set of candidate variables it is clear that selecting them by hand can be very cumbersome, especially if forecasts for many different predictands, stations, and lead times are required.

Thus, algorithms to automatically select the most important variables are highly desirable. The following subsection introduces a boosting algorithm that can be employed for this purpose.

2.2. Non-homogeneous boosting

This subsection introduces an alternative algorithm to the BFGS optimization to estimate the coefficients β and γ. This algorithm is based on boosting and can automatically select the most important predictor variables. Like other optimization algorithms, boosting finds the minimum of the loss function (e.g., the negative log-likelihood; Equation 4) iteratively but in each step it only updates the coefficient of the variable that improves the current fit most. Thus, if it is stopped early and not run until convergence only the most important variables have non-zero coefficients.

In the following we assume the predictand (y) and each predictor variables (x_j, z_k) to have zero mean and unit variance. We use standardized anomalies (see following section for details) to achieve this. Alternatively, one could subtract the mean and divide the standard deviation of each variable. Then the non-homogeneous boosting algorithm can be described by:

1. Initialize coefficients:
$$\beta = \mathbf{0}, \gamma = \mathbf{0} \tag{5}$$

2. Iterate *mstop* times:

   (a) Compute negative partial derivatives of $L(\mu, \sigma)$ with respect to $\mu = \mathbf{x}^\top \beta$ and $\sigma = \mathbf{z}^\top \gamma$:
   $$r = -\frac{\partial L(\mu, \sigma)}{\partial \mu} \qquad s = -\frac{\partial L(\mu, \sigma)}{\partial \sigma} \tag{6}$$

   (b) Find the predictor variable x_j with the highest correlation to r, and z_k with the highest correlation to s:
   $$j^* = \underset{j}{\operatorname{argmax}}\, \rho(x_j, r) \qquad k^* = \underset{k}{\operatorname{argmax}}\, \rho(z_k, s) \tag{7}$$

   (c) tentatively update coefficients:
   $$\beta^* = \beta \qquad \gamma^* = \gamma \tag{8}$$

   $$\beta^*_{j^*} = \beta^*_{j^*} + \nu \rho(x_{j^*}, r) \qquad \gamma^*_{k^*} = \gamma^*_{k^*} + \nu \rho(z_{k^*}, s) \tag{9}$$

   (d) really update the coefficient that improves the current fit most:
   $$\text{if } L(\mathbf{x}^\top \beta^*, \sigma) < L(\mu, \mathbf{z}^\top \gamma^*) \text{ set } \beta = \beta^* \text{ else set } \gamma = \gamma^* \tag{10}$$

where $\mathbf{0}$ are vectors of zeros, *mstop* is a predefined number of boosting iterations, $\rho(x_j, r_m)$ is the sum over the training data of $x_j \times r$, and ν is a predefined step size between 0 and 1. Schmid and Hothorn (2008) showed that the choice of the step size is only of minor importance and we follow their suggestion of $\nu = 0.1$.

If *mstop* is selected to be very large, the estimated coefficients β and γ approximate the maximum likelihood estimates from the model in the previous subsection. However, by choosing a smaller *mstop*, overfitting can be prevented with unimportant variables having zero coefficients. In order to get the best predictive performance an appropriate *mstop* has to be found. This is achieved by optimizing the cross-validated log-likelihood. The data is split in

e.g., 10 parts and for each part, predictions are computed from non-homogeneous boosting models that were fitted on the remaining 9 parts. This is done for different *mstop* from 1 to a rather high $mstop_{max}$ resulting in $mstop_{max}$ different predictions for each event in the data set. *mstop* is then set to the *mstop* with the smallest negative log-likelihood sum over all events.

In addition to automatically selecting the most important input variables, boosting also regularizes the non-zero coefficients, i.e., the coefficients are shrunk compared to their maximum-likelihood values. Hastie *et al.* (2013) showed that this regularization is similar to that of the absolute shrinkage and selection operator (LASSO; Tibshirani 1994) and also helps to reduce overfitting, especially for highly correlated input variables.

In the following we investigate non-homogeneous Gaussian boosting (*NGB*) in a case study and compare its performance with that of *NGR* with only the ensemble forecast of the predictand variable as input. To assess the influence of the regularization in boosting, we also compare a further NGR model with the *subset* of input variables that were selected by boosting.

3. Data

This section describes the data that is used for the case study in the following section. We considered minimum and maximum temperatures at the five central European SYNOP stations Wien Schwechat (48.110N, 16.570E), Innsbruck Airport (47.260N, 11.357E), Berlin Tegel (52.566N, 13.311E), Leipzig Halle (51.436N, 12.241E), and Zürich Kloten (47.480N, 8.536E). Minimum temperatures are for periods between 18UTC and 06UTC, and maximum temperatures between 06UTC and 18UTC.

As numerical predictions we employed the 51 member ensemble predictions from the ECMWF. In addition to the direct forecast of minimum and maximum temperatures we used various predictions for different parameters (e.g., temperatures, wind, precipitation) from the surface level and pressure levels at 1000, 850, 700, and 500 hPa. The regarded 12 hour time windows (18–06UTC or 06–18UTC) span several (3-hourly) time steps of the ECMWF model. For accumulated quantities (e.g., precipitation) we simply employed the accumulated values over the regarded 12 hour time window. For other quantities (e.g., temperatures) we computed means, maxima, and minima over the regarded time windows for each parameter and member respectively.

Subsequently ensemble means and log-standard deviations were derived. The logarithm of the ensemble standard deviations is used to be consistent with the log-scale that is modeled in Equation 3. Zero standard deviations sometimes occur for variables with a limited range such as precipitation. These variables are almost never selected by our models but to avoid infinite numbers we set zero standard deviations to a very small value (0.0001).

For each accumulated parameter this results in two variables (ensemble mean and log-standard deviation) and for each other parameter in six variables (ensemble means and log-standard deviations for 12-hourly means, minima, and maxima). In the following, these variables are labeled according to following rule: *parameter_aggregation_ensemble-statistic*, e.g., **t2m_dmax_mean** is the ensemble **mean** of **d**aily (12 hourly) **max**imum **t**emperature ensemble forecasts at **2m** above ground.

In addition to the ensemble predictions from the numerical weather forecasting model, the last

observed minimum or maximum temperature is used as potential predictor variable. Overall 335 input variables are available to the NGB model.

We regarded lead times from 1 to 5 days (30 to 138 hours) and use data from January 2011 to January 2016 (approx. 1700 days).

Clearly, many variables such as temperatures, have strongly pronounced seasonal patterns that probably affect the statistical properties of forecasts and observations. To only use training data that is representative for the current season, many studies use moving training windows of a certain number of days preceding the forecast date (e.g., Gneiting *et al.* 2005; Thorarinsdottir and Gneiting 2010; Scheuerer and Büermann 2014). While this approach allows the model to adapt quickly to seasonal changes it disregards large parts of available data.

To allow larger training data sets, we used standardized anomalies to remove seasonal patterns. For these standardized anomalies, first seasonally varying climatological means and standard deviations were derived for the predictand and all input variables. Therefore a non-homogeneous regression model (Equations 1 to 3) was fitted with y the respective parameter (predictand or input variable) and $x = z = (1, \sin(2\pi d/365), \cos(2\pi d/365))^\top$. Standardized anomalies are then easily derived for parameter a by:

$$\frac{a - m_a}{s_a} \tag{11}$$

where m_a and s_a are the climatological mean (location) and standard deviation (scale) derived from the non-homogeneous regression model. As an example, Figure 1 shows that the standardized anomalies of maximum temperatures in Wien Schwechat have no pronounced seasonal cycle anymore so that the entire dataset can be used for training.

Note that when anomalies are employed, location predictions $\hat{\mu}$ have to be transformed back by:

$$m_y + \hat{\mu} s_y \tag{12}$$

and scale predictions $\hat{\sigma}$ by:

$$\hat{\sigma} s_y \tag{13}$$

4. Results

This section assesses the boosting algorithm on the data described in the previous section. To illustrates the boosting optimization, Figure 2 shows a typical evolution of coefficients. Since the input variables all have unit variance their coefficient values can be directly compared and indicate their relevance. After all coefficients being zero in the beginning, the daily mean maximum temperature ensemble mean (tmax2m_dmean_mean) is the first variable that gets a non-zero coefficients which indicates that it explains the observations best. With an increasing value of the corresponding coefficient, more and more of the variance in the observations is explained so that the intercept for the log-scale decreases. After approximately 20 iterations the ensemble standard deviation of daily maximum evaporation (ske_dmax_sd) enters with a negative coefficient for the log-scale. Few steps later the daily maximum 2m temperature ensemble mean (t2m_dmax_mean) is added to the equation for the location. Further selected variables are the daily minimum soil temperature ensemble mean (stl1_dmin_mean) and the

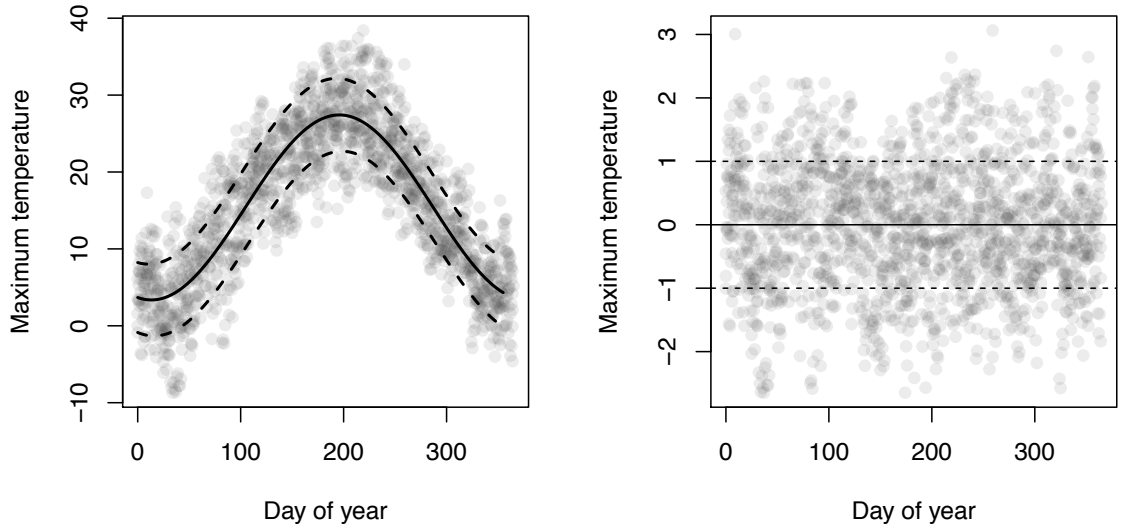

Figure 1: Left: Observed maximum temperatures (gray circles) in Wien Schwechat, fitted climatological mean (solid line), and climatological mean ± one climatological standard deviation (dashed lines). Right: Corresponding standardized anomalies.

700 hPa daily mean vorticity ensemble standard deviation (not labeled) for the log-scale and the daily minimum 1000 hPa temperature ensemble mean (not labeled) for the location. Further variables enter the regression equations later, but are not considered because the optimum cross validation stopping iteration is already found at 31.

Figure 3 shows the boosting coefficients from the cross validation stopping iteration at different lead times for maximum temperature forecasts in Wien Schwechat. Additionally, dashed lines show the NGR coefficients. As already indicated in Figure 2, the daily maximum maximum temperature ensemble forecast, that would be the direct predictor, is neither important for the location nor for the log-scale. However, it is highly correlated (correlation coefficients > 0.9) to e.g., the daily mean maximum temperature, the daily maximum 2m temperature or temperatures at 1000 hPa, so that these variables are virtually exchangeable without loosing much information. For the log-scale (Figure 3 bottom), ensemble standard deviations of various variables are selected but also ensemble mean forecasts (e.g., of 1000 hPa divergence d1000_dmax_mean) seem to contain forecast uncertainty information. Interestingly, the NGR coefficient of the ensemble standard deviation in the scale equation is negative for short lead times indicating a negative spread-skill relationship (Wilks 2011).

Figure 4 shows coefficients similar to Figure 3 but for minimum temperatures. The direct predictor, the daily minimum minimum temperature ensemble mean, is clearly the most relevant variable over all lead times unlike for maximum temperatures. However, various other variables seem to be more relevant for the log-scale equation, many also with negative coefficients. Note that for Wien Schwechat (Figures 3 and 4) boosting selects relatively few

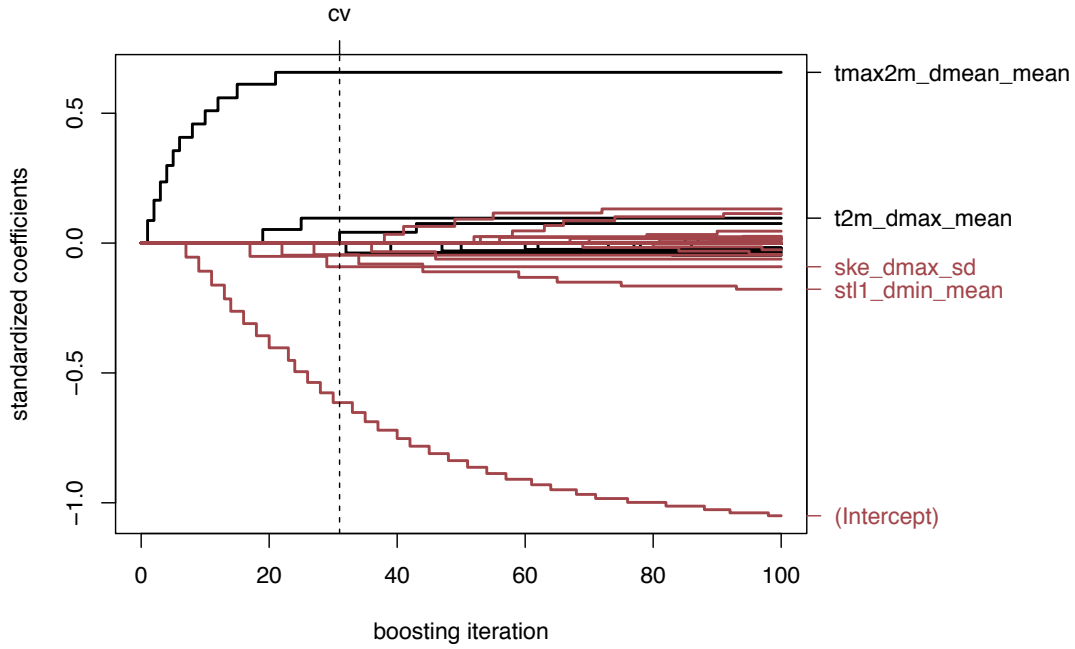

Figure 2: Paths of boosting coefficients for a +66 hours maximum temperature forecasts at Wien Schwechat. Coefficient paths for the location are shown as black lines and for the log-scale as red lines. The optimum stopping iteration according to cross validation (cv) is shown as dashed vertical line. The most important coefficients are labeled (see text).

variables. Many more variables are selected for some of the other stations (not shown).

Figures 2 to 4 show that boosting selects a meteorologically reasonable set of variables. In the following, we investigate how the increased number of input variables improves the forecast performance. To obtain independent training and test data, 10-fold cross validation is used again: For each station and lead time the data is split into 10 parts and for each part performance measures (squared errors, CRPS or PITs) are computed for models that were trained on the 9 remaining parts. The effective training data length is thus 9/10 of the full data set length (approximately 1550 days). To estimate the sampling distribution of average squared errors and CRPS we computed means of 250 bootstrap samples.

Figure 5 shows the root mean squared error (RMSE) of the location forecasts (μ in Equation 2) of NGB, NGR, and the subset NGR, which is an NGR with the non-zero coefficients from boosting as input. For the two stations, Wien Schwechat and Innsbruck Airport, the RMSE of the minimum temperature forecast is always smaller for boosting than for NGR. As already indicated in Figure 4, NGR and boosting differ only slightly for Wien minimum temperature forecasts. In contrast the differences are much larger for Innsbruck. In addition to selecting the most important variables, boosting also regularizes or shrinks the coefficients. The subset model uses the same variables as boosting but does not regularize their coefficients which results in very similar RMSE. The RMSE of the other stations and maximum temperatures

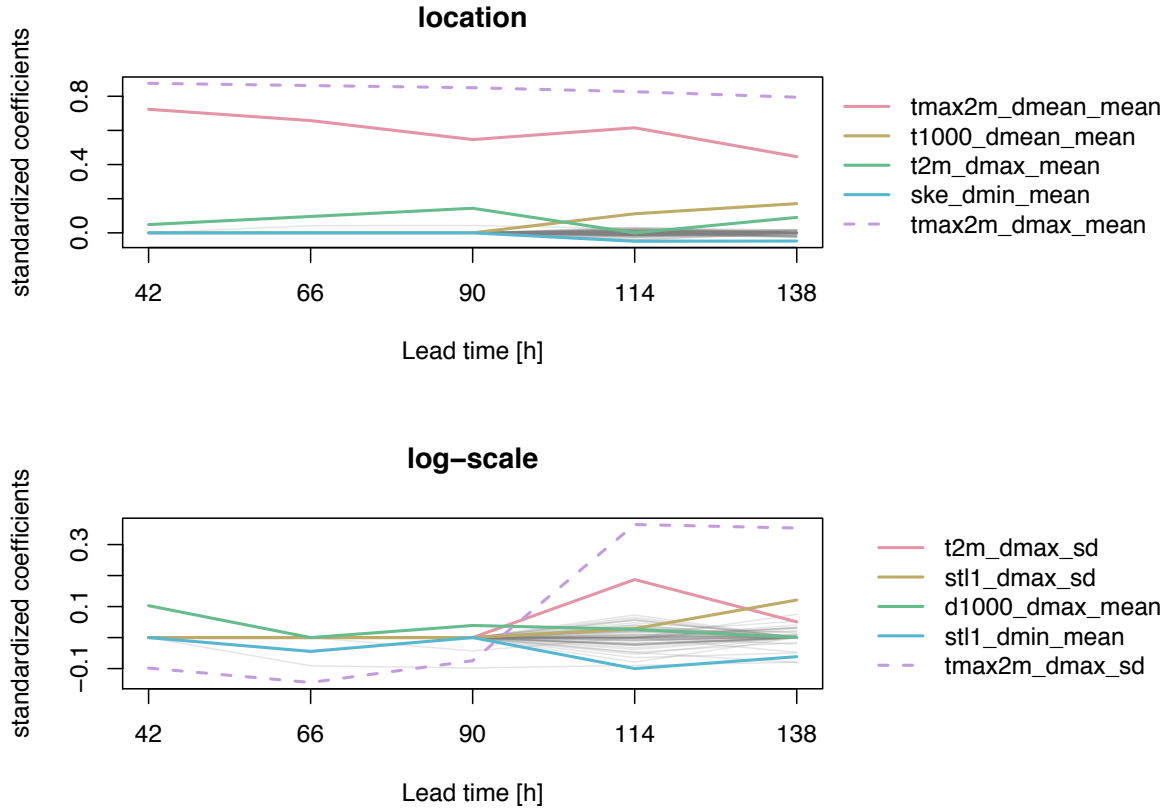

Figure 3: Standardized coefficients from non-homogeneous boosting for Wien Schwechat maximum temperature for the location (top) and the log-scale deviation (bottom) and different lead times. The intercepts are not shown and the most important coefficients are shown in colors. The optimum stopping iteration was found by cross validation.

look very similar to that of Wien Schwechat and Innsbruck Airport minimum temperatures and are therefore not shown.

While the RMSE shows the deterministic performance, we employ the continuous ranked probability score (CRPS; Hersbach 2000) to measure the probabilistic quality of the forecasts. Gneiting *et al.* (2005) provides a closed form for normal predicitve distributions

$$CRPS = \sigma \left\{ \frac{y - \hat{\mu}}{\hat{\sigma}} \left[2\Phi \left(\frac{y - \hat{\mu}}{\hat{\sigma}} \right) - 1 \right] + 2\phi \left(\frac{y - \hat{\mu}}{\hat{\sigma}} \right) - \frac{1}{\sqrt{\pi}} \right\} \tag{14}$$

where $\Phi()$ and $\phi()$ are the normal cumulative distribution function and probability density function respectively, y is the observation and $\hat{\mu}$ and $\hat{\sigma}$ are the predicted location and scale. Since we are mainly interested in improvements of boosting over NGR, Figure 6 shows the continuous ranked probability *skill* score (CRPSS) relative to NGR

$$CRPSS = 1 - \frac{\overline{CRPS}}{\overline{CRPS}_{NGR}} \tag{15}$$

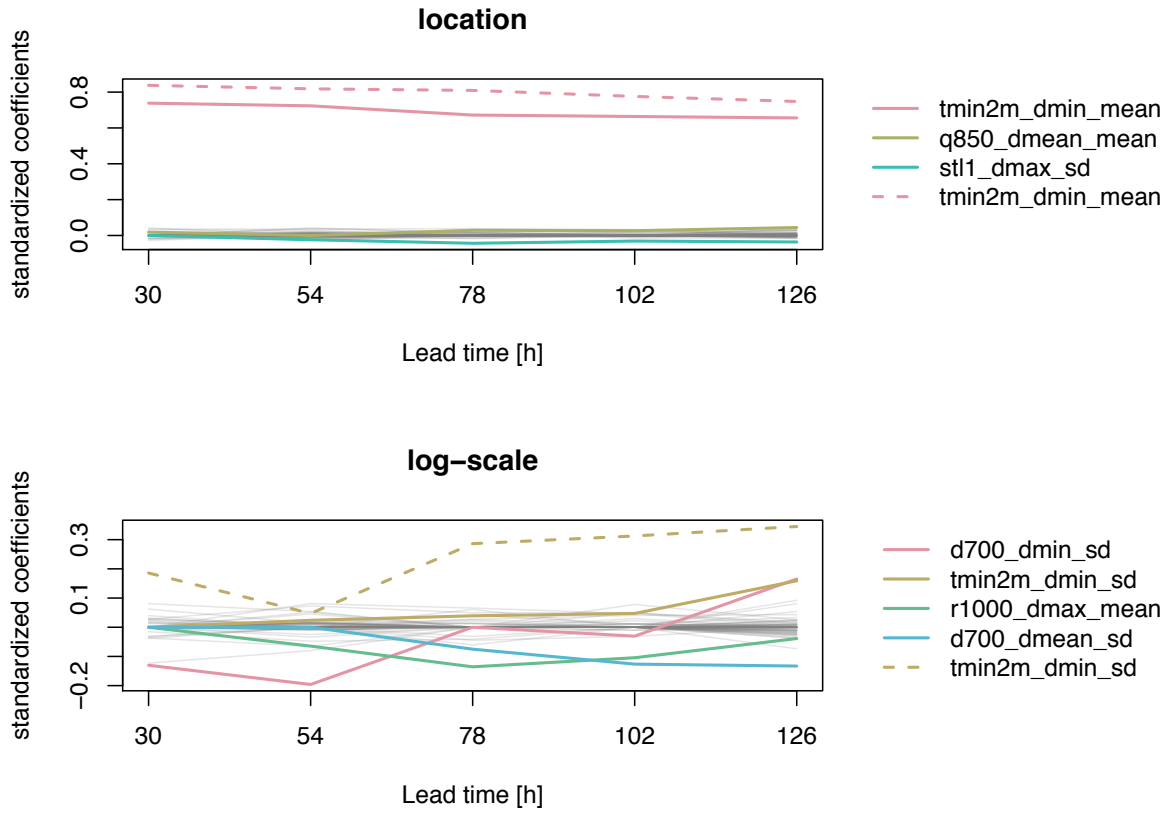

Figure 4: Same as Figure 3 but for Wien Schwechat minimum temperature.

where $\overline{CRPS}$ is the respective average CRPS and $\overline{CRPS}_{NGR}$ is the average CRPS of NGR. For both, minimum and maximum temperature forecasts, NGB performs clearly better than NGR over all lead times where for longer lead times this advantage is less pronounced. Different to the RMSE the regularization in boosting slightly improves the forecast performance compared to the subset model.

To assess the reliability of the forecasts, Figure 7 shows probability integral transform (PIT) histograms (Wilks 2011) of NGB and NGR. Both forecast methods seem to produce predictive distributions with too light left and too heavy right tails, indicating that actually a non-symmetric distribution would better fit the data. However, the flatter PIT histogram of NGB indicates that using more variables partly compensates for this problem and increases the reliability.

Finally, Figure 8 shows the CRPSS for different training data lengths. For shorter training data lengths the number of selected input variables decreases but is still proportionally high compared to the training data length. In the subset model this leads to overfitting that clearly deteriorates the predictive performance. In contrast, NGB regularizes the coefficients to largely prevent overfitting so that, except for very short training data lengths, it outperforms NGR.

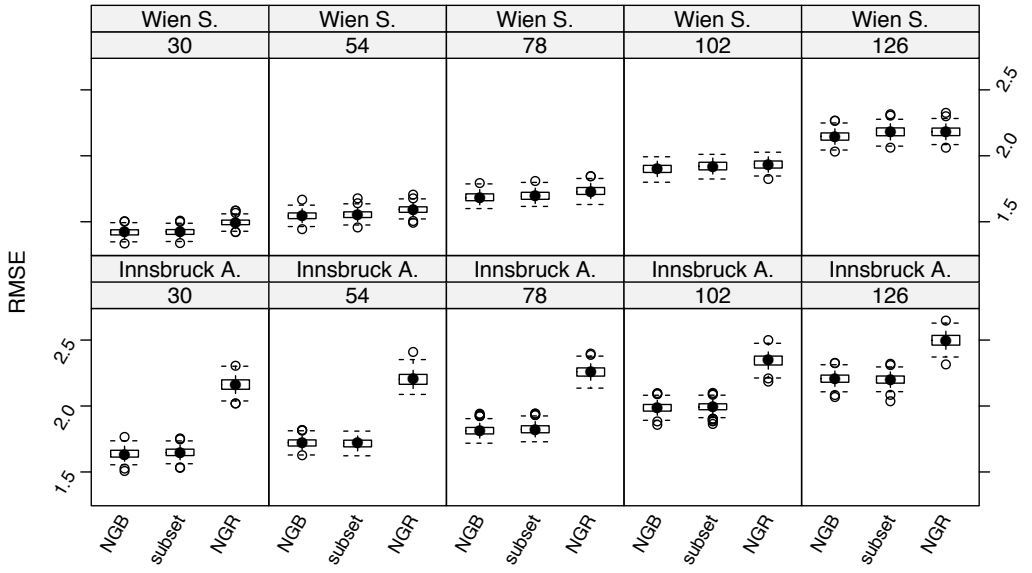

Figure 5: Root mean squared error for Wien Schwechat (top) and Innsbruck Airport (bottom) minimum temperature forecasts for different lead times and models. The solid circles mark the medians and the boxes the interquartile range of the 250 RMSE values from bootstrapping. The whiskers show the most extreme 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.

5. Summary and conclusion

Non-homogeneous regression can easily be extended to use further predictor variables in addition to ensemble forecasts of the predictand variable. However, to avoid overfitting that can deteriorate the predictive performance, predictor variables have to be selected carefully.

In this paper we presented a boosting algorithm to estimate the regression coefficients that can be used for automatic variable selection. In addition to variable selection this algorithm also regularizes or shrinks the regression coefficients to further prevent overfitting. A case study for minimum and maximum temperatures at five central European statios showed clear improvements in the predictive performance compared to a non-homogeneous regression model with only ensemble mean and standard deviation of the predictand variable as input.

In our case study we employed a large set of different ensemble predictions from ECMWF (approx. 100) at surface and several pressure levels. We aggregated these predictions over the regarded time windows and computed ensemble means and log-standard deviations. Additionally, we also used the last available observations as potential predictor variable. Clearly there are many more potential input variables that we have not included; e.g., current observations of other variables or from neighboring weather stations, deterministic predictions or ensemble predictions from other centers, transformations of all of these variables (e.g., logarithm, roots, or powers), etc. Including some of these would probably further improve the forecasts.

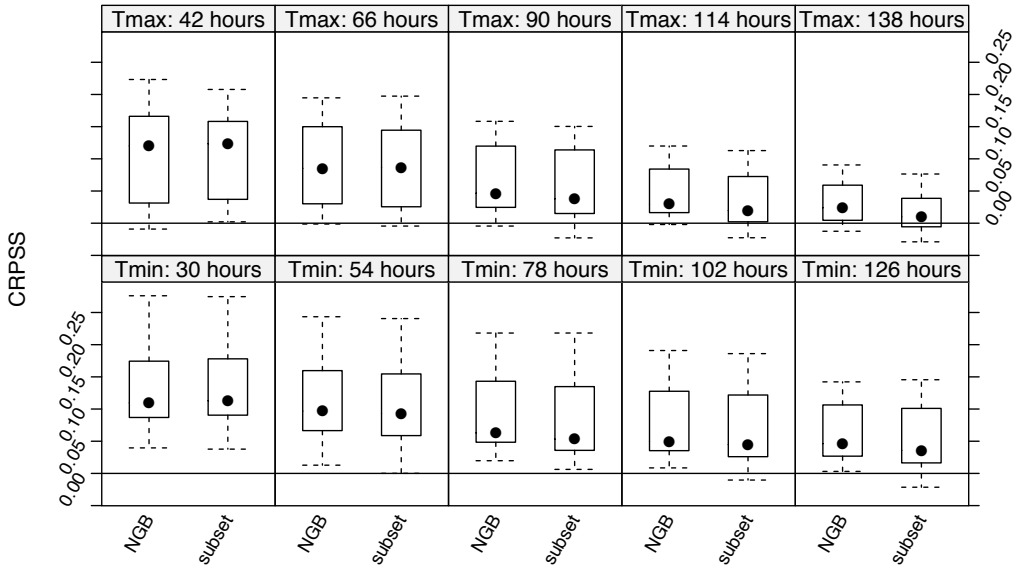

Figure 6: Continuous ranked probability skill score (CRPSS) relative to NGR of maximum (top) and minimum (bottom) temperature forecasts aggregated over 5 stations. Circles, boxes and whiskers have the same meaning as in Figure 5.

In this paper we assumed minimum and maximum temperatures to follow normal distributions. However, the PIT histograms indicate that the conditional distribution of maximum and minimum temperatures given the ensemble forecast is not perfectly symmetric so that using a different asymmetric distribution could improve the forecast performance. Other distributions might also be required for predictions of other non-normally distributed variables such as precipitation or wind speed. Although we presented boosting for normal distributed predictive distributions, all concepts can be easily transfered to other distributions as well. Similarly, also other differentiable loss functions, such as the CRPS, could be employed instead of the negative log-likelihood.

Variable selection is clearly not new in the statistical post-processing literature. Glahn and Lowry (1972) already recognized the importance of variable selection for deterministic model output statistics and proposed to use stepwise selection. However, except Broecker (2010) who proposed lasso regularization for logistic regression and Wahl (2015) who used lasso penalization for quantile regression, automatic variable selection has not been used in the *ensemble* post-processing literature so far.

Non-homogeneous boosting is an easy to implement extension of the popular non-homogeneous regression to automatically select the most relevant input variables of possibly very large sets of candidates. To facilitate the implementation and adaption to other problems we provide all our algorithms in the software package `crch` (Messner *et al.* 2016) for the open source software `R` (R Core Team 2015).

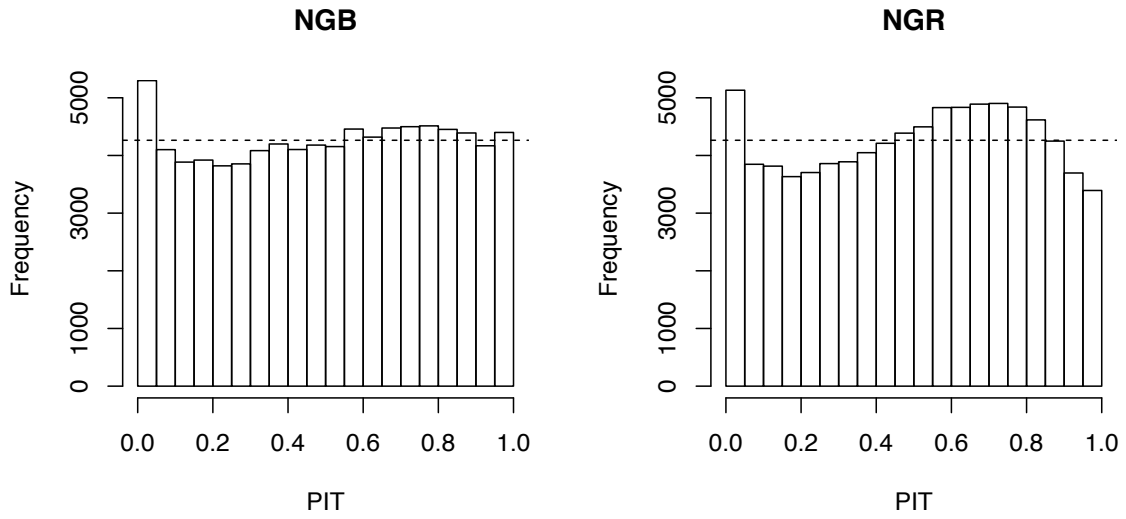

Figure 7: Probability Integral Transform (PIT) histogram aggregated over 5 stations, lead times, and predictands for non-homogeneous boosting (left) and non-homogeneous regression (right). Perfect PIT uniformity is indicated by horizontal dashed lines.

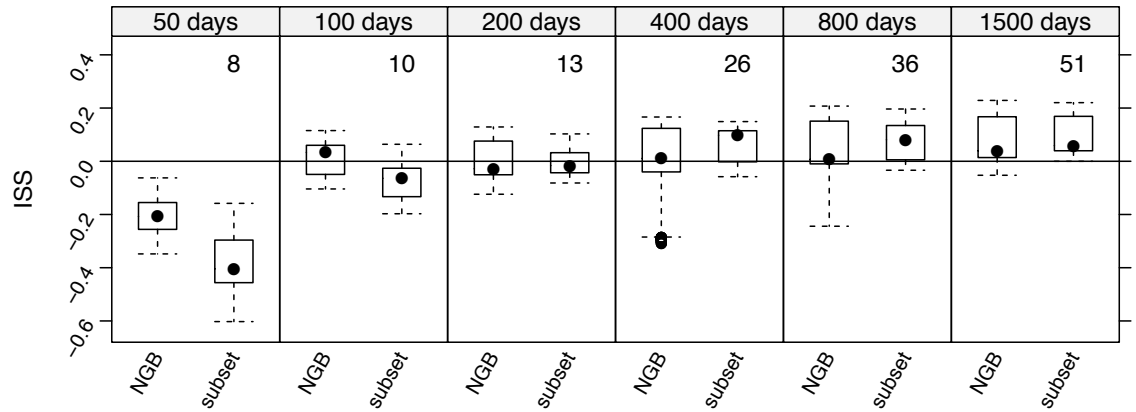

Figure 8: Continuous ranked probability skill score (CRPSS) relative to NGR for 42 hours maximum temperature forecasts and different training data lengths aggregated over 5 stations. The respective median numbers of selected input variables are shown at the top of each panel.

6. Acknowledgments

This study was supported by the Austrian Science Fund (FWF) TRP 290-N26. Data from the ECMWF forecasting system were obtained from the ECMWF Data Server.

References

Broecker J (2010). "Regularized Logistic Models for Probabilistic Forecasting and Diagnostics." *Mon. Wea. Rev.*, **138**(2), 592–604. doi:10.1175/2009MWR3126.1.

Bühlmann P, Hothorn T (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting." *Statistical Science*, **22**(4), 477–505. doi:10.1214/07-STS242.

Bühlmann P, Yu B (2003). "Boosting With the L2 Loss: Regression and Classification." *Journal of the American Statistical Association*, **98**, 324–339.

Feldmann K, Scheuerer M, Thorarinsdottir TL (2015). "Spatial Postprocessing of Ensemble Forecasts for Temperature Using Nonhomogeneous Gaussian Regression." *Mon. Wea. Rev.*, **143**(3), 955–971. doi:10.1175/MWR-D-14-00210.1.

Freund Y, Schapire RE (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences*, **55**(1), 119 – 139. doi:10.1006/jcss.1997.1504.

Friedman J, Hastie T, Tibshirani R (2000). "Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors)." *The Annals of Statistics*, **28**(2), 337–407. doi:10.1214/aos/1016218223.

Glahn H, Lowry D (1972). "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting." *J. Appl. Meteor.*, **11**(8), 1203–1211. doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005). "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation." *Mon. Wea. Rev.*, **133**(5), 1098–1118. doi:10.1175/MWR2904.1.

Hamill TM, Whitaker JS, Wei X (2004). "Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts." *Mon. Wea. Rev.*, **132**(6), 1434–1447. doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hastie T, Tibshirani R, Friedman J (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition. Springer.

Hersbach H (2000). "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems." *Weather and Forecasting*, **15**(5), 559–570. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Messner JW, Mayr GJ, Wilks DS, Zeileis A (2014a). "Extending Extended Logistic Regression: Extended vs. Separate vs. Ordered vs. Censored." *Mon. Wea. Rev.*, **142**, 3003–3014. doi:10.1175/MWR-D-13-00355.1.

Messner JW, Mayr GJ, Zeileis A (2016). "Heteroscedastic Censored and Truncated Regression with crch." *The R Journal*, p. to appear.

Messner JW, Zeileis A, Mayr GJ, Wilks DS (2014b). "Heteroscedastic Extended Logistic Regression for Post-Processing of Ensemble Guidance." *Mon. Wea. Rev.*, **142**, 448–456. doi:10.1175/MWR-D-13-00271.1.

Pinson P (2012). "Adaptive Calibration of (u,v)-Wind Ensemble Forecasts." *Quart. J. Roy. Meteor. Soc.*, **138**(666), 1273–1284. doi:10.1002/qj.1873.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005). "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Mon. Wea. Rev.*, **133**(5), 1155–1174. doi:10.1175/MWR2906.1.

Roulston MS, Smith LA (2003). "Combining Dynamical and Statistical Ensembles." *Tellus A*, **55**(1), 16–30. doi:10.1034/j.1600-0870.2003.201378.x.

Schefzik R, Thorarinsdottir TL, Gneiting T (2013). "Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling." *Statistical Science*, **28**(4), 616–640. doi:10.1214/13-STS443.

Scheuerer M (2014). "Probabilistic Quantitative Precipitation Forecasting using Ensemble Model Output Statistics." *Quart. J. Roy. Meteor. Soc.*, **140**(680), 1086–1096. doi:10.1002/qj.2183.

Scheuerer M, Büermann L (2014). "Spatially Adaptive Post-Processing of Ensemble Forecasts for Temperature." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**(3), 405–422. doi:10.1111/rssc.12040.

Scheuerer M, Hamill TM (2015). "Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions." *Mon. Wea. Rev.*, **143**(11), 4578–4596. doi:10.1175/MWR-D-15-0061.1.

Schmid M, Hothorn T (2008). "Boosting Additive Models Using Component-Wise P-Splines." *Computational Statistics & Data Analysis*, **53**(2), 298 – 311. doi:10.1016/j.csda.2008.09.009.

Schuhen N, Thorarinsdottir TL, Gneiting T (2012). "Ensemble Model Output Statistics for Wind Vectors." *Mon. Wea. Rev.*, **140**(10), 3204–3219. doi:10.1175/MWR-D-12-00028.1.

Thorarinsdottir TL, Gneiting T (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. doi:10.1111/j.1467-985X.2009.00616.x.

Tibshirani R (1994). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Wahl S (2015). *Uncertainty in Mesoscale Numerical Weather Prediction: Probabilistic Forecasting of Precipitation*. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Wilks DS (2009). "Extending Logistic Regression to Provide Full-Probability-Distribution MOS Forecasts." *Meteor. Appl.*, **368**(March), 361–368. doi:10.1002/met.134.

Wilks DS (2011). *Statistical Methods in the Atmospheric Sciences*. 3 edition. Academic Press.

Wilson LJ, Vallé M (2002). "The Canadian Updateable Model Output Statistics (UMOS) System: Design and Development Tests." *Weather and Forecasting*, **17**(2), 206–222. doi:10.1175/1520-0434(2002)017<0206:TCUMOS>2.0.CO;2.

Affiliation:

Jakob W. Messner, Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstraße 15
6020 Innsbruck, Austria
E-mail: Jakob.Messner@uibk.ac.at, Achim.Zeileis@uibk.ac.at

Georg J. Mayr
Institute of Atmospheric and Cryospheric Sciences
Faculty of Economics and Statistics
Universität Innsbruck
Innrain 52
6020 Innsbruck, Austria
E-mail: Georg.Mayr@uibk.ac.at

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2016-04 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Non-homogeneous boosting for predictor selection in ensemble post-processing

2016-03 **Dietmar Fehr, Matthias Sutter:** Gossip and the efficiency of interactions

2016-02 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Rankings and risk-taking in the finance industry

2016-01 **Sibylle Puntscher, Janette Walde, Gottfried Tappeiner:** Do methodical traps lead to wrong development strategies for welfare? A multilevel approach considering heterogeneity across industrialized and developing countries

2015-16 **Niall Flynn, Christopher Kah, Rudolf Kerschbamer:** Vickrey Auction vs BDM: Difference in bidding behaviour and the impact of other-regarding motives

2015-15 **Christopher Kah, Markus Walzl:** Stochastic stability in a learning dynamic with best response to noisy play

2015-14 **Matthias Siller, Christoph Hauser, Janette Walde, Gottfried Tappeiner:** Measuring regional innovation in one dimension: More lost than gained?

2015-13 **Christoph Hauser, Gottfried Tappeiner, Janette Walde:** The roots of regional trust

2015-12 **Christoph Hauser:** Effects of employee social capital on wage satisfaction, job satisfaction and organizational commitment

2015-11 **Thomas Stöckl:** Dishonest or professional behavior? Can we tell? A comment on: Cohn et al. 2014, Nature 516, 86-89, "Business culture and dishonesty in the banking industry"

2015-10 **Marjolein Fokkema, Niels Smits, Achim Zeileis, Torsten Hothorn, Henk Kelderman:** Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees

2015-09 **Martin Halla, Gerald Pruckner, Thomas Schober:** The cost-effectiveness of developmental screenings: Evidence from a nationwide programme

2015-08 **Lorenz B. Fischer, Michael Pfaffermayr:** The more the merrier? Migration and convergence among European regions

2015-07 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Cooperation and discrimination within and across language borders: Evidence from children in a bilingual city

2015-07 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Cooperation and discrimination within and across language borders: Evidence from children in a bilingual city

2015-06 **Martin Geiger, Wolfgang Luhan, Johann Scharler:** When do Fiscal Consolidations Lead to Consumption Booms? Lessons from a Laboratory Experiment

2015-05 **Alice Sanwald, Engelbert Theurl:** Out-of-pocket payments in the Austrian healthcare system - a distributional analysis

2015-04 **Rudolf Kerschbamer, Matthias Sutter, Uwe Dulleck:** How social preferences shape incentives in (experimental) markets for credence goods *forthcoming in Economic Journal*

2015-03 **Kenneth Harttgen, Stefan Lang, Judith Santer:** Multilevel modelling of child mortality in Africa

2015-02 **Helene Roth, Stefan Lang, Helga Wagner:** Random intercept selection in structured additive regression models

2015-01 **Alice Sanwald, Engelbert Theurl:** Out-of-pocket expenditures for pharmaceuticals: Lessons from the Austrian household budget survey

University of Innsbruck

Working Papers in Economics and Statistics

2016-04

Jakob W. Messner, Georg J. Mayr, Achim Zeileis

Non-homogeneous boosting for predictor selection in ensemble post-processing

Abstract
Non-homogeneous regression is often used to statistically post-process ensemble forecasts. Usually only ensemble forecasts of the predictand variable are used as input but other potentially useful information sources are ignored. Although it is straightforward to add further input variables, overfitting can easily deteriorate the forecast performance for increasing numbers of input variables. This paper proposes a boosting algorithm to estimate the regression coefficients while automatically selecting the most relevant input variables by restricting the coefficients of less important variables to zero. A case study with ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) shows that this approach effectively selects important input variables to clearly improve minimum and maximum temperature predictions at 5 central European stations.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)