Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and the Environment

Improving GLAMEPS wind speed forecasts by statistical postprocessing

Dorien Lugt

De Bilt, 2013 | Intern Rapport; IR-2013-03

Royal Netherlands
Meteorological Institute
*Ministry of Infrastructure
and the Environment*

TU Delft Delft University of Technology

Improving GLAMEPS Wind Speed Forecasts by Statistical Postprocessing

This report is written as part of the internship of

Dorien Lugt

at the
Royal Netherlands Meteorological Institute
(KNMI)

Department of Weather Research

May 1, 2013 – July 12, 2013

Supervisors:

Maurice Schmeits
Kees Kok

Preface

This report is written as part of the internship of Dorien Lugt at the Royal Netherlands Meteorological Institute (KNMI) at the department of Weather Research. The internship was under supervision of Maurice Schmeits and Kees Kok (KNMI) and took place between May 1 and July 12, 2013. At the time of this internship Dorien Lugt was an Applied Mathematics undergraduate student at Delft University of Technology. The aim of the project was to investigate whether the GLAMEPS wind speed forecasts could be improved by statistical postprocessing.

Abstract

The Grand Limited Area Model Ensemble Prediction System (GLAMEPS) is a short range multi-model ensemble system, a combination of the subensembles from the ALADIN, HIRLAM STRACO and Kain-Fritsch and ECMWF models. In this report it is shown that the (gridbox average) GLAMEPS 10 meter wind speed forecasts can be improved significantly by statistical postprocessing when verified against station observations. Results show that both reliability and resolution can be improved for wind speeds up to 16 m/s. Special attention has been paid to the improvement for higher wind speeds. Furthermore, experiments indicated that use of the mean of the control runs of the four models instead of the ensemble mean do not result in losses in Brier skill scores for forecasts 18, 30 and 42 hours ahead. Postprocessing of the complete ensemble performed better than postprocessing each of the four subensembles separately. These conclusions were derived from data from November and December 2011 for stations all over Europe and need to be confirmed for datasets of longer periods and outputs of the current GLAMEPS version.

1. Introduction

Weather forecasts are derived from the output of numerical weather prediction (NWP) models that describe the processes in the atmosphere. The output of NWP models is deterministic, whereas probabilistic predictions are preferred for several reasons. Because of the complexity of the processes in the atmosphere, NWP models do not describe the atmospheric processes completely and perfectly. Furthermore, the estimated initial state of the model differs from the real state of the atmosphere. This can lead to incorrect deterministic forecasts even when predicting only a few days ahead. Additional information about the uncertainty of a deterministic prediction, as given by a probabilistic prediction, is desirable.

The state-of-the-art method of producing probabilistic weather predictions is the use of ensembles. An ensemble of initial states reflecting the uncertainty of the initial state is created and each member is used as the initial state for integration with the NWP model. This results in an ensemble of deterministic predictions, which can be combined into a probabilistic prediction.

Statistical postprocessing of the ensemble output is used to correct for systematic errors in the forecasts and to include the effect of local conditions that are absent because of the scale of the grid of the numerical models. Statistical postprocessing is performed on a dataset containing both the model output and the observations. Using regression - in this study extended logistic regression (ELR) and logistic regression (LR) - it can be determined which combination of the model output parameters appear to have the highest predictive value; these parameters can be used for future forecasts.

The ensemble system used in this project is GLAMEPS, *Grand Limited Area Model Ensemble Prediction System*. GLAMEPS is a system for short range probabilistic numerical weather predictions on European scale. GLAMEPS is part of the cooperation between two European short-range model consortia: Aire Limitée Adaption dynamique Developpement International (ALADIN) and High Resolution Limited Area Modeling (HIRLAM). It consists of four subensembles, ALADIN, HIRLAM STRACO, HIRLAM Kain-Fritsch and ECMWF, each having 13 members (Iversen et al., 2011). In the two HIRLAM subensembles different convection schemes have been used: STRACO (Sass, 2002) and Kain-Fritsch (Kain and Fritsch, 1990). GLAMEPS has been running since 2010.

The aim of this study is to improve the forecast probability distribution of (high) wind speeds. Different predictors for the regression have been investigated as well as other aspects such as the training area and the use of subsamples of the total number of ensemble members. Also, logistic regression has been compared to extended logistic regression. In section 2 of this report the dataset and the regression method are described. In section 3 the results are shown. Conclusions, discussion and recommendations for future research can be found in section 4 of this report.

2. Data and Method

In this section two regression methods (LR and ELR) and a number of verification metrics are described.
For this project part of the GLAMEPS output parameters from July until December 2011 were available. Because in the winter months wind speeds are higher, it was decided to use the month November for training and the month December for verification. This is a short period, which was partly compensated for by using 2249 European stations with observations available for the same period.
The predictand in this study was the wind speed at a height of 10 meters. The only GLAMEPS output parameters available in the dataset were the 10 meter wind speed, the 2 meter temperature and the 12 hours accumulated precipitation. The 10 meter wind speed and the latitude, longitude and altitude of the stations are useful predictors for wind speed forecasts. Observed wind speeds of more than 50 m/s are considered to be unlikely high and were therefore not used in the regression and verification. Unless mentioned otherwise, experiments were performed for 18 hours lead time.

2.1 Logistic regression and extended logistic regression

Since the predictand is binary, i.e. the wind speed is *lower* than a certain threshold or not[1], a natural choice for the regression method is logistic regression, which has the form

$$p = \frac{e^{f(x)}}{1 + e^{f(x)}} \tag{1}$$

Here f(x) is a linear function of predictors (Wilks, 2006). Using logistic regression the coefficients have to be derived for each threshold separately. Wilks (2009) described an extension of logistic regression, referred to as extended logistic regression (ELR) from here on, that overcomes this problem. Using extended logistic regression the full forecast probability distribution is obtained so that forecasts can be derived for thresholds that were not used in the training. Another advantage of extended logistic regression is the consistency among the resulting probabilities for different thresholds. Extended logistic regression takes the form

$$p(q) = \frac{e^{f(x)+g(q)}}{1 + e^{f(x)+g(q)}} \tag{2}$$

where f(x) is defined as in the logistic regression and q is the threshold. So:

$$p(q) = \Pr[windspeed \le q] \tag{3}$$

Wilks (2009) empirically determined that the function $g(q) = b\sqrt{q}$ performed better than $g(q) = bq$, with *b* a logistic regression coefficient to be determined together with the coefficients of f(x). Even though Wilks focused on precipitation instead of wind speed in his article, it was decided to set $g(q) = b\sqrt{q}$ in this study as well, without further investigations of other nondecreasing functions g(q).

1) Note that we have taken probabilities less than the threshold instead of exceedence probabilities. This choice does not affect the results.

2.2 Brier Scores and Brier Skill Scores

For the verification several performance measures are known. The measures used in this project are the Brier score (BS) with its decomposition and the Brier skill score (BSS) (Wilks, 2006). The BS is the mean of the squared difference between the forecast probabilities and the observations:

$$BS = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - o_k\right)^2 \tag{4}$$

In this equation y_k are the forecast probabilities, between 0 and 1, and o_k are the observations, having value 0 or 1. The BS has a value between zero and one, for perfect forecasts the BS is zero and a worse forecast gets a higher BS. A decomposition of the BS can be made so that three factors that play a role in the performance of a probabilistic forecast system are separated, namely the reliability, the resolution and the uncertainty:

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i\left(y_i - \bar{o}_i\right)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i\left(\bar{o}_i - \bar{o}\right)^2 + \bar{o}\left(1-\bar{o}\right) \tag{5}$$

= "Reliability" - "Resolution" + "Uncertainty

Here I is the number of allowable (or binned) forecast values.
The reliability is a measure for how close the binned forecast probabilities are to the observed conditional relative frequencies. For perfectly reliable forecasts the forecasted probabilities (y_i) are exactly equal to the conditional relative frequencies ($\bar{o}_i$), thus the difference is zero and so is the reliability score. A higher reliability score implies less reliability. Resolution reflects how much the conditional observed frequencies ($\bar{o}_i$) differ from the sample climatological average ($\bar{o}$). A resolution of zero means that the forecasts are exactly the climatological average; the highest possible resolution is equal to the uncertainty score. The uncertainty score reflects the inherent uncertainty of the event, i.e. zero for low uncertainty and ¼ for maximum uncertainty. Reliability and resolution may be improved by statistical postprocessing, but the uncertainty cannot be influenced.
The Brier skill score (BSS) can be used to compare the BS of two models, for example the postprocessed GLAMEPS and a reference forecast model:

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{6}$$

The BSS is the relative improvement of, e.g., the new statistical model over the reference model: if the BSS is zero the new model has the same BS as the reference model, if it is negative the BS of the new model is worse. If the BSS is between zero and one, the new method has a better BS than the reference method. The maximum value of the BSS is 1, or 100% when expressed as a percentage.

3. Results

Several possibilities for improvement of the GLAMEPS forecast probability distribution of wind speed using statistical postprocessing have been investigated. First different potential predictors have been compared. Next the sensitivity of the performance of the postprocessing for the thresholds used for training was investigated. The model configuration derived this way was also verified on a limited area and compared with a model that was both trained and verified on a limited area. Also, logistic regression has been compared to extended logistic regression. Finally, the use of the mean of the control runs of the four subensembles (instead of the ensemble mean) and the use of the mean per subensemble (instead of the ensemble mean) have been tested. In all experiments November 2011 was used as the training month and December 2011 as the verification month, so independent data were used for the verification. In this section the results of these experiments are discussed. Unless mentioned otherwise, about 100000 training cases varying in station, time and threshold were used, which is about half of the total number of cases.

<u>3.1 The BSS of the raw GLAMEPS forecasts</u>

All BSSs in this report were computed using the BS of the GLAMEPS forecasts before postprocessing, i.e. raw GLAMEPS forecasts, as a reference. However, Figure 1 shows the BSSs per threshold of the raw GLAMEPS forecasts with the climatological probabilities as a reference. The BSSs of the raw GLAMEPS probabilities and the climatological average were averaged over all stations and not calculated per station. Only thresholds up to and including 16 m/s were included in the results, because the number of higher observed wind speeds in the dataset was too low to guarantee sufficient significance of the results.

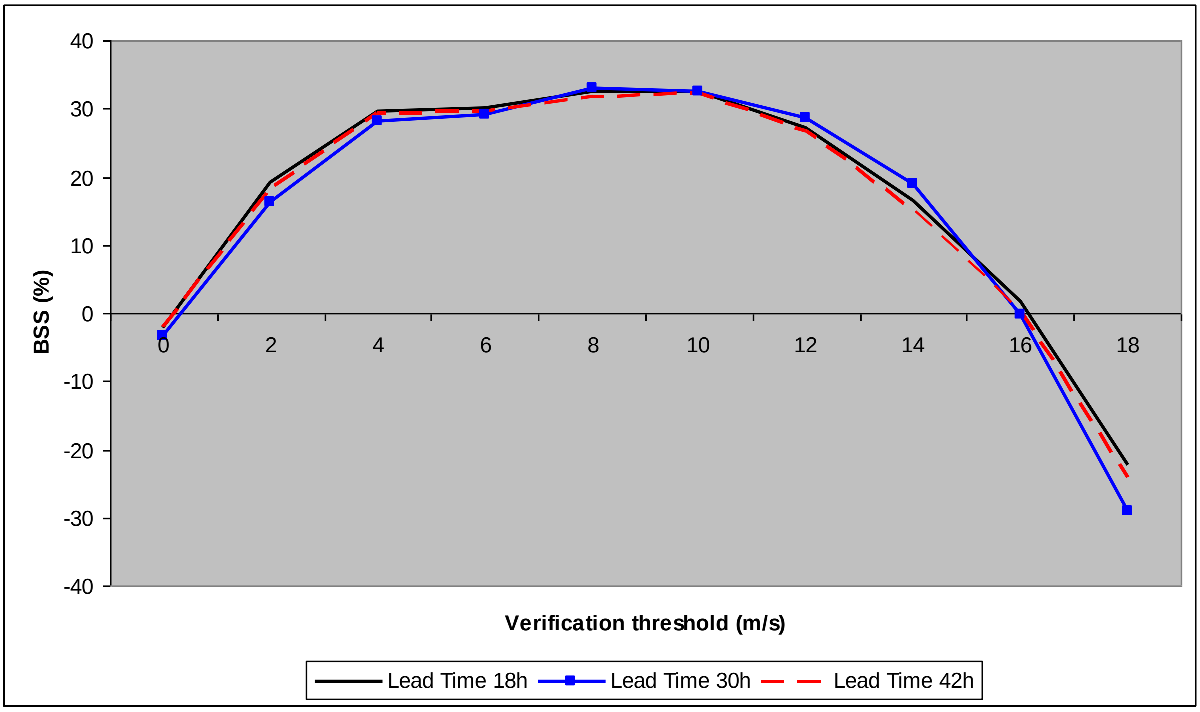


Figure 1. BSS of the raw GLAMEPS forecasts with respect to the climatological probabilities

In Figure 1 the BSSs of the raw GLAMEPS forecasts are positive for thresholds of 2-14 m/s and are almost equal for three different lead times (18 hours, 30 hours and 42 hours). This may be due to the averaging of the climatological probabilities over all stations. It is expected that the BSSs decrease for longer lead times when calculated per station.

<u>3.2 Tests with different predictors</u>

As mentioned before, the dataset available for this study was limited in the sense that not all GLAMEPS output parameters were available. Therefore, other potential predictors than those that were tested exist. The GLAMEPS wind speed at 10 meter height was available and the ensemble mean of this parameter turned out to be a good predictor. Since wind speeds are not normally distributed, better results were obtained when using the mean of the square root of each ensemble member’s wind speed output instead of the mean of those wind speeds itself. From here on this will be referred to as the ensemble mean predictor. Also, the longitude, latitude and the altitude of the meteorological stations appeared to be good predictors and their performance could be further improved by using a smoothing spline. In this project the thin plate regression spline as described in Wood (2003) was used. The standard deviation of the ensemble did not have added value over the ensemble mean predictor.
Another potential predictor in the dataset was the difference between the altitude of the stations and the altitude of the common model grid. The NWP models are grid-based, so that if a station is not exactly located on the grid, its altitude is estimated by averaging over the nearest grid points. Especially in regions with higher topography, the estimated altitude can differ significantly from the

real elevation level of the stations. Experiments with the difference between the real and the common model altitude of the stations as a predictor have been performed in order to investigate whether this would reduce any systematic errors in the forecasts. It turned out that this predictor resulted in higher BSS than the station altitude for lower thresholds, but lower BSS for higher thresholds. Since in this study the focus lies on improvement of the forecasts for higher wind speeds, it was decided not to include this predictor in the postprocessing.
Concluding, from the parameters available in the dataset the ensemble mean predictor and the spline function of the longitude, latitude and altitude of a station turned out the be good predictors and were included in the regression model, together with the square root of the thresholds for ELR.

3.3 The sensitivity for the thresholds used for training

It was suggested by Wilks (2009) that improvements on the performance of extended logistic regression could be obtained by tuning the thresholds and the number of thresholds on which the regression model is being trained. Using the square root of the thresholds, the ensemble mean predictor and the spline function of the longitude, latitude and the elevation as predictors, the performance for different combinations of training thresholds has been compared. For every selected combination of thresholds in the training set the BSS was calculated on the independent data for all even thresholds from 0 m/s until 16 m/s. It turned out that the scores for higher wind speeds were, not surprisingly, better for combinations of higher thresholds and likewise for lower wind speeds.
Of all the tested combinations the best performing combinations are displayed in Figure 2. In addition to these combinations the combination 0, 1, 5, 10 and 15 m/s is added to the figure for comparison.

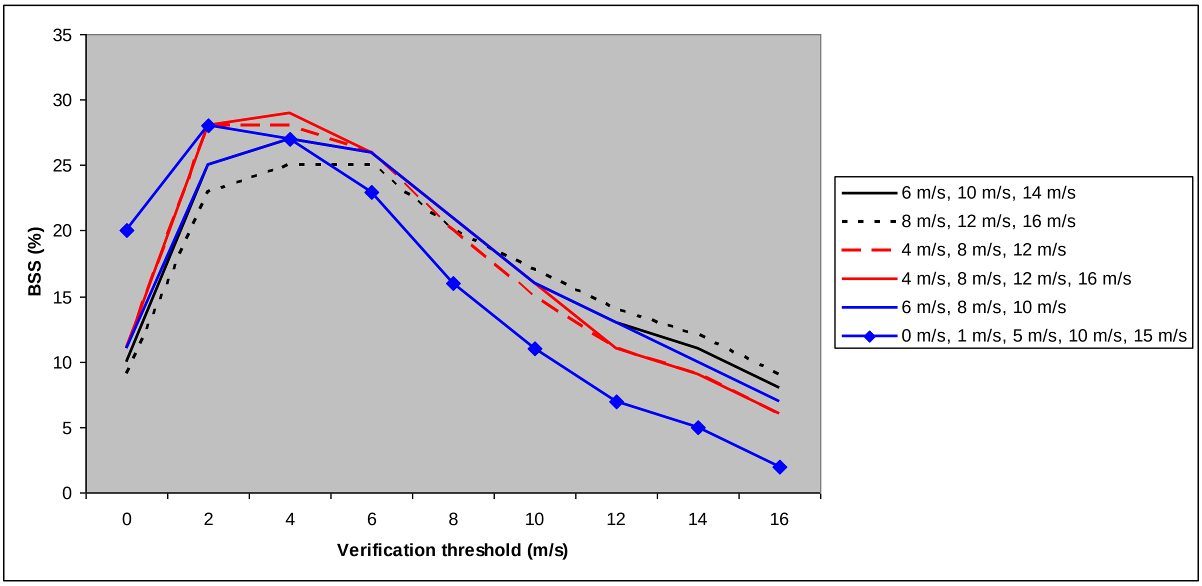


Figure 2. BSS for ELR models using different sets of training thresholds

The choice for the best combination is somewhat subjective. It was decided to use the combination of 8, 12 and 16 m/s, since this combination performed best for the higher thresholds. For the lower thresholds the BSSs were lower than for the other combinations but improvement with respect to the raw GLAMEPS was still high, up to 25% for lower thresholds.

<u>3.4 Improvement in terms of Reliability and Resolution</u>

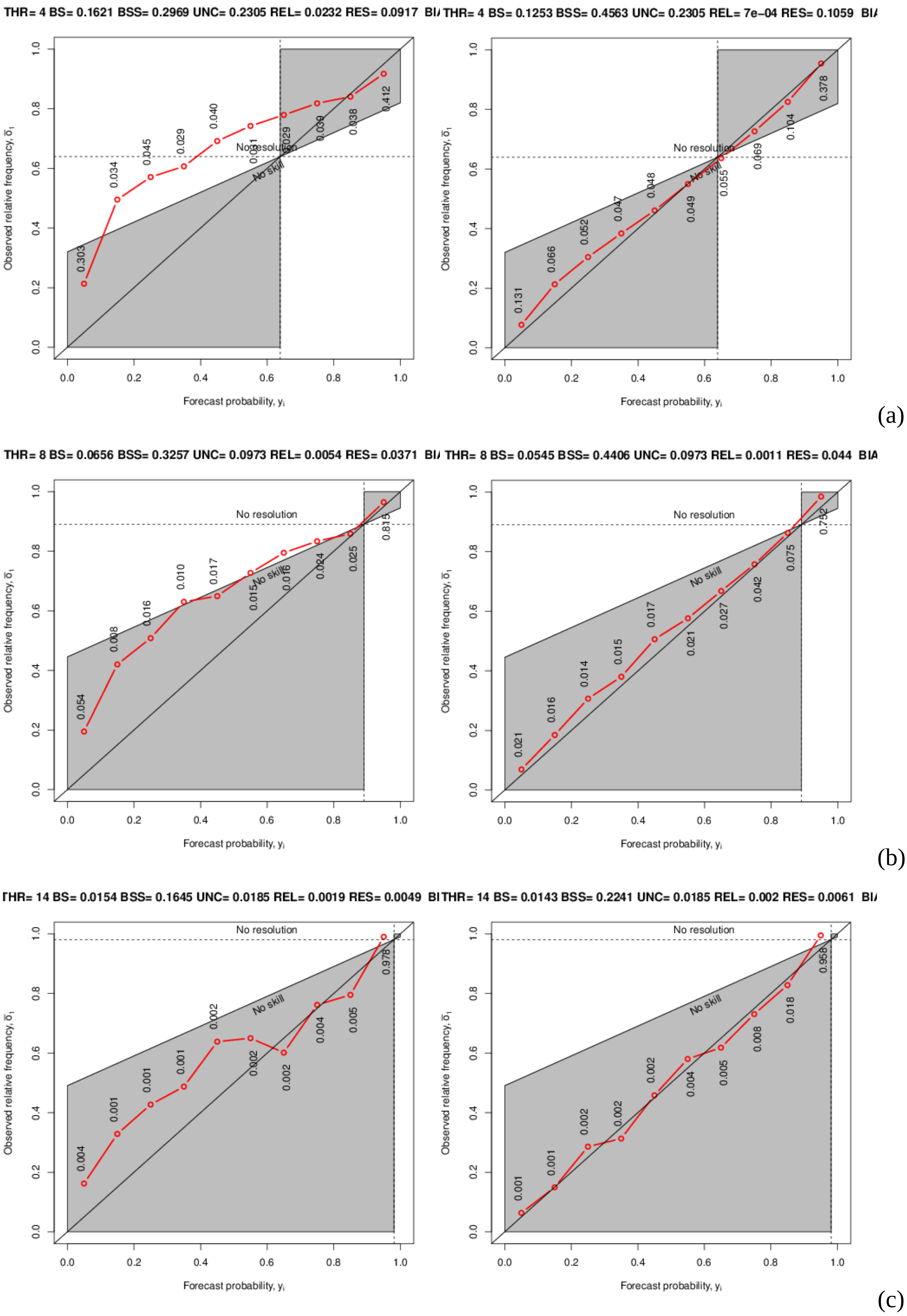


Figure 3. Reliability diagram for the raw GLAMEPS forecasts (left) and after postprocessing (right), a) for a threshold of 4 m/s, b) for a threshold of 8 m/s and c) for a threshold of 14 m/s. The frequencies of the forecasted probabilities are given for each circle on the red line.

It has been shown that significant improvements in terms of the BS can be realized by statistical postprocessing. In section 2.2 the decomposition of the BS was described. Reliability diagrams (Wilks, 2006) give insight in the improvement in terms of reliability and resolution. In Figure 3 the reliability diagrams for the raw GLAMEPS forecasts (left panels) and for the postprocessed model (right panels) are shown for thresholds of 4, 8 and 14 m/s (Figs a, b and c respectively) .
As explained before, reliability is a measure for the difference between the forecasted probabilities and the conditional observed relative frequencies. A reliability score of zero means that the forecasted probabilities are equal to the observed relative frequencies, so the forecasts are perfectly reliable. In the reliability diagram this means that the red line lies exactly on the diagonal.
For the 4 m/s threshold (Figure 3a) the postprocessing of GLAMEPS clearly has improved the reliability, which is also represented by the decrease in reliability score from 0.0232 to 0.0007. The resolution, as explained before, is a measure of the difference between the conditional observed frequencies and the sample climatological average. It is clear that the frequencies of the forecast probabilities close to 0 and close to 1 have increased, so the resolution of the postprocessed GLAMEPS was higher than that of the raw GLAMEPS. This is also represented by an increase in the resolution score from 0.0917 to 0.1059. So, for the prediction of the threshold of 4 m/s both reliability and resolution have improved.
Figure 3b shows the reliability diagrams of both methods for the prediction threshold of 8 m/s. Here an improvement in both reliability (from 0.0054 to 0.0011) and resolution (from 0.0371 to 0.0044) is apparent as well.
For the more extreme threshold of 14 m/s (Figure 3c) the resolution score improved from 0.0049 to 0.0061. The figure shows clearly that the reliability has improved, as the red line lies closer to the diagonal for almost all forecasted probabilities. However, this is not represented by the reliability score that has increased from 0.0019 to 0.002, suggesting that the reliability for the raw GLAMEPS was higher. This is because for more than 95% of the forecasts the forecasted probability was in the highest bin and this point lies slightly further from the diagonal for the postprocessed GLAMEPS (right panel, Figure 3c) than for the raw GLAMEPS (left panel, Figure 3c). However, even though the reliability score has not improved, the reliability over the whole range of forecasted probabilities has improved.
Similar results were obtained for all other thresholds between 0 m/s and 16 m/s. It can thus be concluded that statistical postprocessing does not only improve the raw GLAMEPS forecasts in terms of BSS, but also in terms of both reliability and resolution. These results were obtained for training and verification on stations all over Europe; experiments on a limited domain have been performed and are described in section 3.6.

3.5 Logistic Regression versus Extended Logistic Regression

In section 2.1 the differences between logistic regression and extended logistic regression are described. Extended logistic regression has the advantage that the complete probability density function can be derived at once so that there are no inconsistencies in forecasts between the thresholds as can occur when using logistic regression. In this paragraph the performance of logistic regression and extended logistic regression is compared. First, for the logistic regression the raw GLAMEPS probabilities, i.e. the percentage of ensemble members that did not exceed the threshold, have been tested as a predictor. It turned out that this predictor had no added value over the ensemble mean predictor and the spline function. Therefore it was not included in the model for logistic regression. The ELR model as used before has been compared to the logistic regression model with the same predictors, except for the square root of the thresholds that is an extra predictor for extended logistic regression. The results are shown in Figure 4. This figure shows that for thresholds above 6 m/s there was (almost) no difference in BSS and that for lower thresholds the logistic regression model performed better than the extended logistic regression model. This can probably be explained by the fact that the ELR model was trained on higher thresholds. Since the focus is on higher wind speeds in this study and taking the advantages of the ELR method into account, it was decided to continue working with ELR.

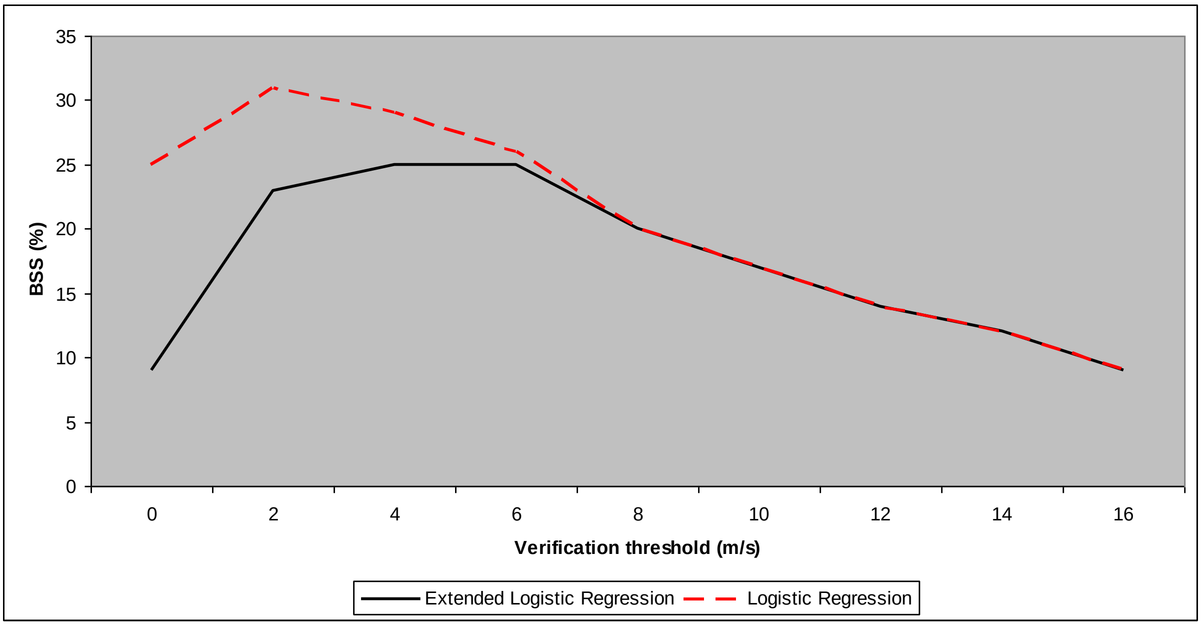


Figure 4. BSS of regression models using logistic regression compared with an ELR model

3.6 Training and Verification on a Limited Domain

The results derived in earlier sections were derived for stations all over Europe. In this section the performance on a limited domain (Figure 5) is described. The domain contains all 284 stations with longitude between 2° and 15°E, latitude between 49° and 55°N, and an altitude less than 500 meters above sea level.

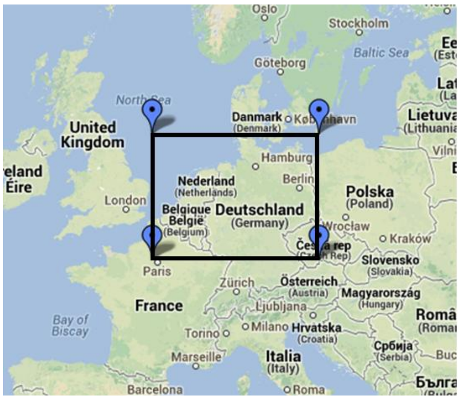


Figure 5. Map showing the limited domain with longitude between 2° and 15°E and latitude between 49° and 55°N

Figure 6 shows the BSSs of the regression model that was trained on stations from all over Europe and verified on the limited domain, but also the model that was both trained and verified on the limited domain. Since the dataset was limited in time, the limited domain could not be chosen too small, since the number of training cases would be reduced too much. In this case the number of training cases was reduced from 100000 to 25000. In order to make a fair comparison, a model that was trained on stations all over Europe but using only 25000 randomly chosen training cases is included in the figure as well. It is clear that training on a limited domain yielded better BSSs for higher thresholds than the models that were derived from stations all over Europe. The model trained and verified on the limited domain but without the spline function of the longitude, latitude and the altitude of the stations performed better for thresholds up to 12 m/s.

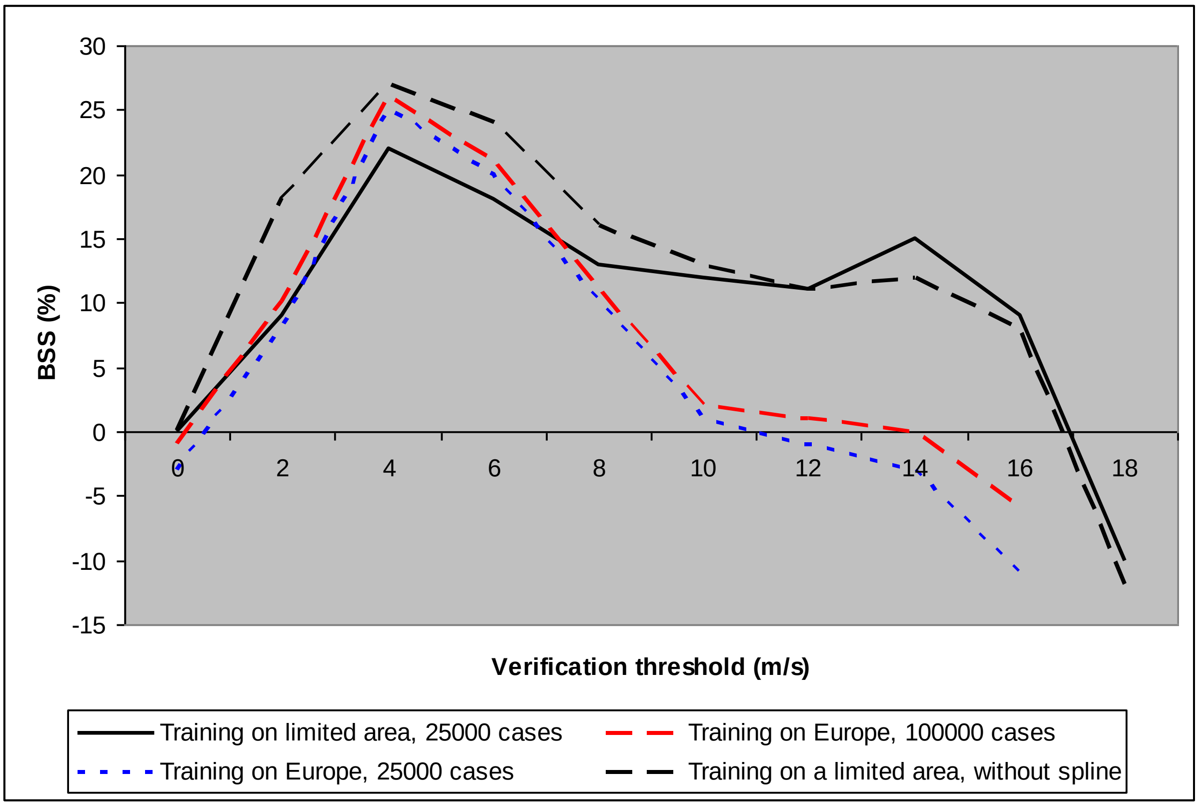
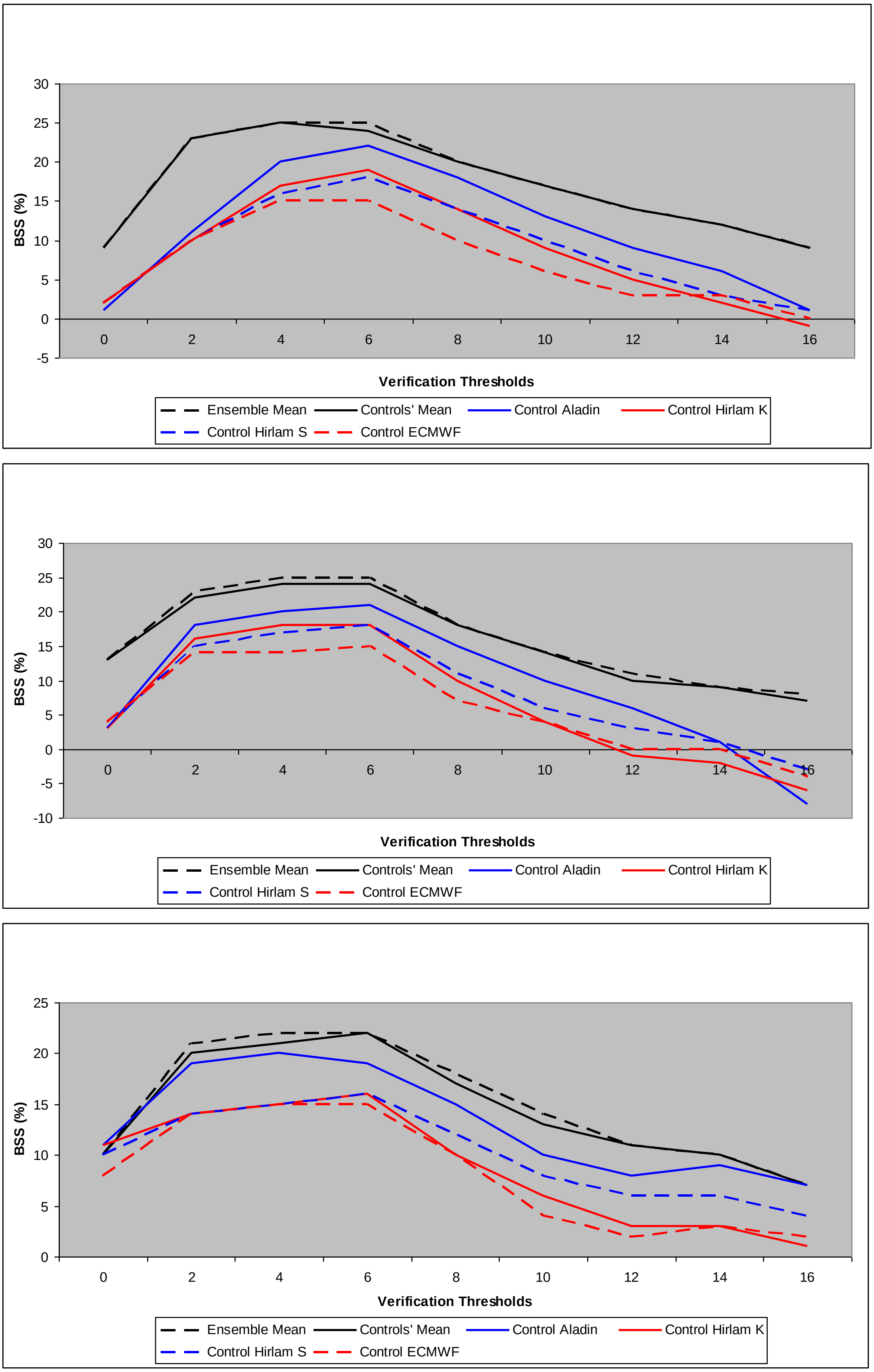


Figure 6. BSS for ELR models verified on a limited area

It can be concluded that significant improvements can be realized by training on a limited domain. The experiments described in section 3.7 however are all executed on the initially chosen domain.

<u>3.7 Controls' mean or subensemble mean instead of ensemble mean</u>

The use of the ensemble mean predictor has been compared to the use of the mean of the control runs of the four subensembles as a predictor. In figure 7 it is shown that the BSSs for these two configurations were (almost) equal for lead times of +18, +30 and +42 hours (Figs 7a, b and c respectively). The results for only one control run as a predictor are shown for all four model configurations as well. It is clear that the use of the mean of the four control runs instead of only one gave better results. This implies that the multi-model character of the GLAMEPS ensemble has advantages over a single model forecast for wind speed, but that for the predictors used in this regression the 48 extra ensemble members did not seem to have much added value over the mean of the four control runs.



(a)

(b)

(c)

Figure 7. BSS of ELR models with the ensemble mean predictor, the controls' mean, the ALADIN control run, the HIRLAM K control run, the HIRLAM S control run and the ECMWF control run as predictors for a lead time of a) 18 hours, b) 30 hours and c) 42 hours.

In figure 8 the use of the mean of each subensemble as a predictor has been compared with the use of the ensemble mean predictor. It is shown that the ensemble mean predictor clearly performed better than each subensemble mean separately. When comparing the four subensembles, the ensemble of ALADIN performed best in this experiment. This was also concluded by Iversen et al. (2011, fig. 10b). Similar results were obtained for lead times of 30 hours and 42 hours (not shown).

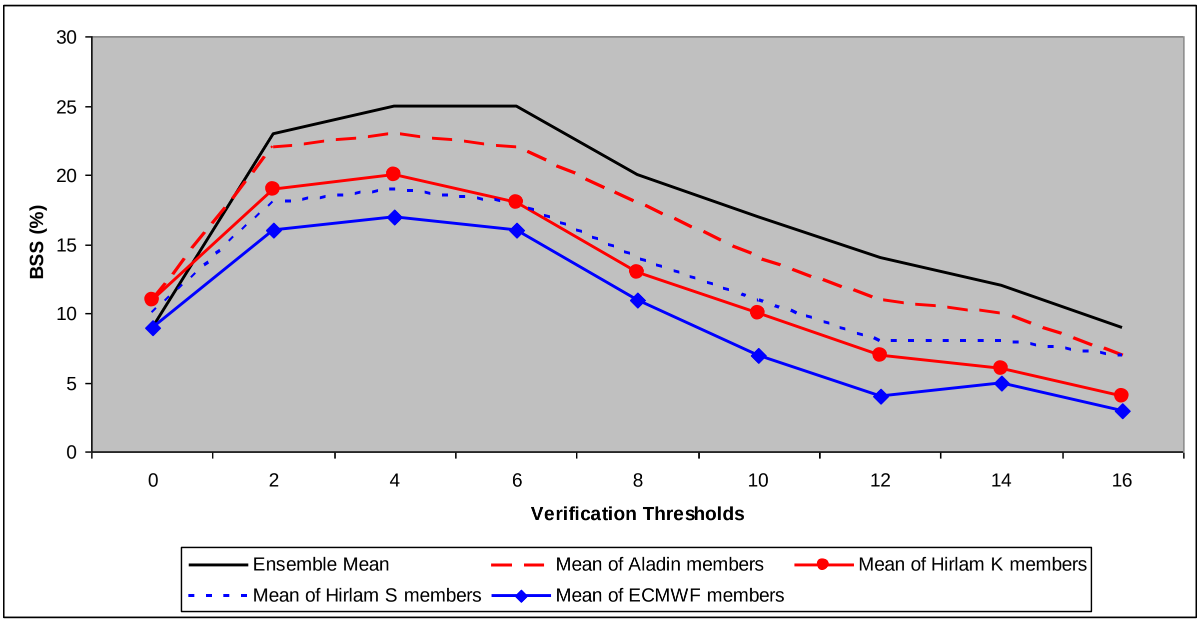


Figure 8. BSS of ELR models with the ensemble mean predictor, the mean of the ALADIN ensemble, the mean of the HIRLAM K ensemble, the mean of the HIRLAM S ensemble and the mean of the ECMWF ensemble

4. Conclusions and discussion

In this report, it has been shown that in deriving local probabilistic wind speed forecasts significant improvements in both the reliability and resolution can be obtained by statistical postprocessing of the raw, i.e. gridbox average, GLAMEPS output for thresholds up to and including 16 m/s. Best predictors from the available dataset were the ensemble mean of the GLAMEPS 10 meter wind speed and a spline function of the longitude, latitude and the altitude of the stations. The ensemble standard deviation and the difference between the real altitude and the model altitude of the stations appeared to have no added value given the other two predictors. Other parameters from the output of GLAMEPS might be worth studying, but they were not available for this project.
The fact that the spline function appeared to be a good predictor suggests that statistically postprocessing GLAMEPS might succesfully be applied also at locations where no observations are available. In a study by De Rooy and Kok (2004) it was shown that the applicability of statistical postprocessing is indeed not limited by the absence of local observations. This was demonstrated on wind speed forecasts using a combination of statistical and physical postprocessing techniques. Splines were not used in that study.
Training the regression model on a limited domain can further improve results for that domain, especially for higher thresholds. On a limited domain, consisting of about 20 percent of all available stations, the postprocessing performed better without the use of the spline of the longitude, latitude and altitude as predictors.
The use of higher thresholds in the training dataset did improve verification results for higher thresholds. For those thresholds the performance of extended logistic regression was approximately equal to the performance of logistic regression, although for lower thresholds logistic regression performed better.
Remarkably, performance of the postprocessed mean of the four control runs was almost equal to the performance of the ensemble mean predictor. Apparently a lot of the probabilitistic information is already contained in the deterministic model. A similar result was found by Kok and Vogelezang

(1999), in which probabilistic temperature forecasts derived from postprocessing the operational deterministic ECMWF model performed better than the “raw” EPS probabilties for much of the 10 day forecast range.
Postprocessing on the GLAMEPS ensemble mean performed better than the use of only one control run or the mean of each subensemble separately. The mean of the ALADIN subensemble performed best of all four subensembles for the prediction of wind speed for lead times of 18, 30 and 42 hours.
These results were obtained for data from November and December 2011. The results need to be confirmed for larger datasets and the current GLAMEPS version.

Acknowledgments

I am grateful to John Bjørner Bremnes for his R program for ELR and for the dataset of GLAMEPS forecasts and observations. I am also grateful for the support of Maurice Schmeits and Kees Kok during this project.

References

De Rooy W.C., and Kok C. J. 2004. A Combined Physical-Statistical Approach for the Downscaling of Model Wind Speed. *Weather and Forecasting* **19**: 485-495.

Iversen T., Deckmyn A., Santos C., Sattler K., Bremnes J. B., Feddersen H., Frogner I. 2011. Evaluation of ‘GLAMEPS’ – a proposed multimodel EPS for short range forecasting. *Tellus* **63A**: 513-530.

Kain J.S., Fritsch J.M. 1990. A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.* **47**: 2784-2802.

Kok C.J. and Vogelezang D.H.P. 1999. Statistical post-processing on EPS. *Joint report on EPS Experts and Medium-Range Users meetings (1999/2000)*: 57 – 64.

Sass, B. H. 2002. A research version of the STRACO cloud scheme. *DMI Tech. Rep. 02-10*. Danish Meteorological Institute, Copenhagen, Denmark, 25 pp. [Available online at http://www.dmi.dk/dmi/index/viden/dmi-publikationer/tekniskerapporter.htm.]

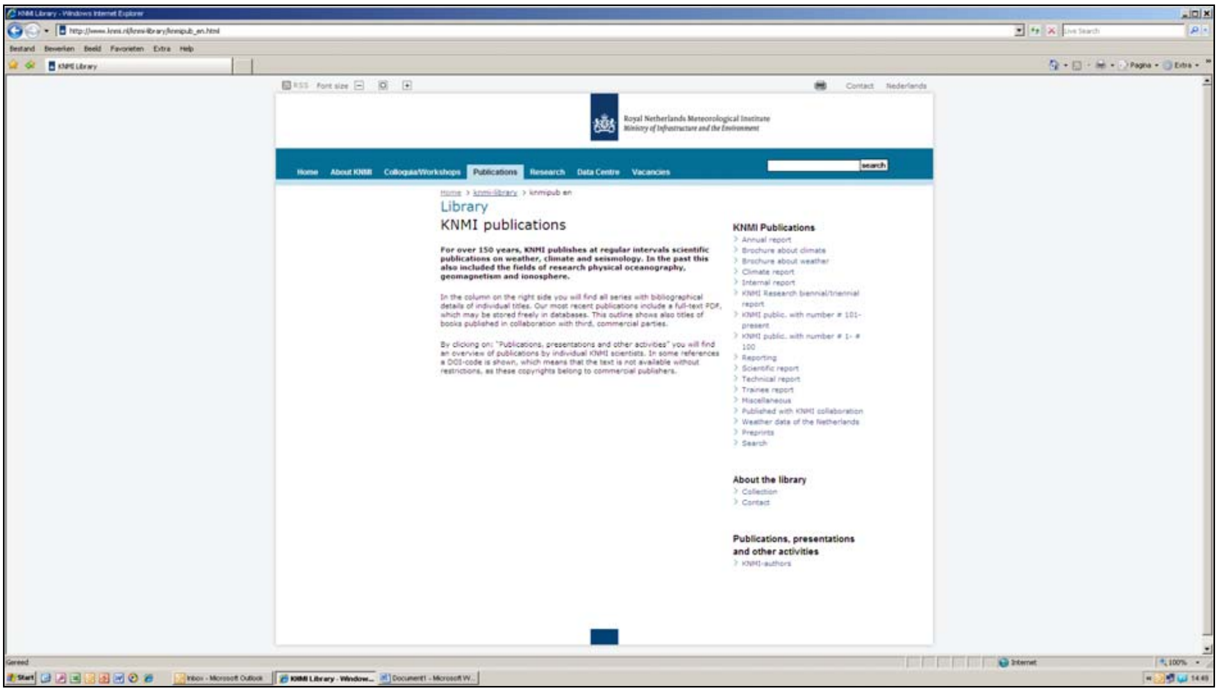
Wilks D. S. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd Edn. Academic Press.

Wilks D.S., 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications* **16**: 361-368.

Wood S. N. 2003. Thin plate regression splines. *J. R. Statist. Soc.* **B65**(1): 95 – 114.

A complete list of all KNMI-publications (1854 – present) can be found on our website

www.knmi.nl/knmi-library/knmipub_en.html



The most recent reports are available as a PDF on this site.