PSO2004/FU5766
Improved wind power prediction

Optimal combined wind power forecasts using exogenous variables

Fannar Örn Thordarson
Henrik Madsen
Henrik Aalborg Nielsen

Technical Report No. 17

October 31, 2007

Informatics and Mathematical Modelling
Technical University of Denmark

Contents

Summary

The aim of combining forecasts is to reduce variation from observed values by composite two or more forecasts, which predict for the same event at the same time. Many methods have developed since the problem was presented, ranging from a method of equal weights to more complex methods, e.g. state space methods. Despite this complexity a linear model of the combination appears to be most acquired where the parameters of the forecasts are summing to one. The parameters, also called weights, are unknown and need to be estimated to get optimal combined forecast. In this report the problem of combining forecasts is addressed by (i) estimating weights by local regression and comparing with recursive least squares and minimum variance methods, which are well known procedures within combining, and (ii) using information from meteorological forecasts to estimate the forecast weights with local regression.

The methods are applied to the Klim wind farm using three WPPT forecasts based on different weather forecasting systems. It is shown how the prediction is improved when the forecasts are combined by using locally fitted linear model and that it outperforms the RLS estimation which is also considered. Furthermore, the meteorological forecasts from *DMI-HIRLAM* are inspected and the air density and the turbulent kinetic energy at pressure level 38 are found to be optimal regressors for locally fitting the weights into of linear combination model.

The results in this report show that using the meteorological information to estimate the weights gives a resonable fit compared to the reference models, which can be elevated by further analysis.

1 Introduction

Where more than one forecast for some event at the same time is available, it can be attractive procedure to combine the forecasts. By combining the independent information included in every individual forecast, more accurate prediction can be accomplished. The application of combining wind power forecasts for certain wind power plant is appealing procedure when several meteorological (MET) forecasts are accessible for the power plant. The MET forecasts are generated to predict for the power production, but different MET forecasts provide different power forecasts. On the market energy is sold in advance but production of wind energy is variable such that a good forecast is needed. With several such forecasts, more accurate forecast can be acquired by combining.

In the following studies the weights are tracked over time for the linear model by considering the recursive least squares method compared with the minimum variance method. The weights are then fitted with local regression.The objective is to attain estimated weights for the combined wind power prediction by conditioning the weights on one or more MET forecasts. The block diagram in Figure 1 shows the flow of combining forecasts and also how the information from MET forecasts are applied to estimate appropriate weights for the combination. By estimating the weights using MET forecasts, external information are added to the combination where the weights do not depend on past data.

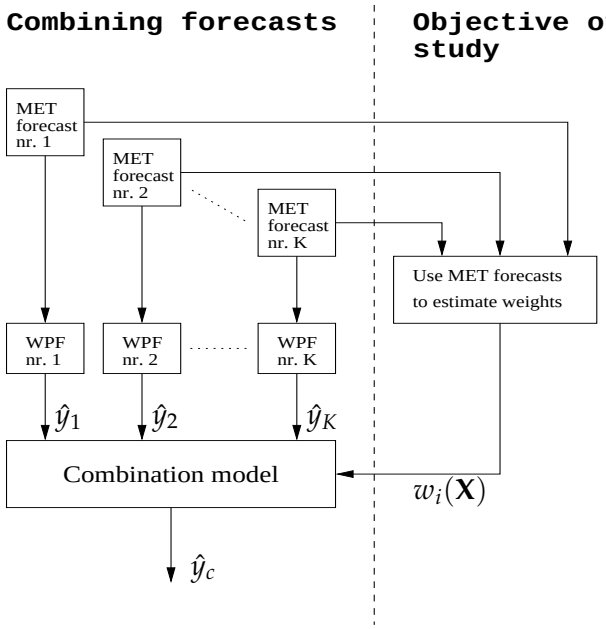


Figure 1: Block diagram of the process of combining wind power forecasts. To the left of the dashed line the flow for combining wind power forecasts (individual forecasts abbreviated WPF in diagram) is described, but the right side shows the black box model for the weights with the MET forecasts as input.

2 Data

The data used in the analysis is twofold, data set including three wind power predictions from WPPT (Wind Power Prediction Tool (Madsen et al., 2005a)) and data set of meteorological forecasts which are used to estimate weights in a combined forecast. Both data sets contain forecasts at time point 00Z, midnight, for every hour over the following 24 hours. There are 7272 data points in the data sets which span the period from February 2nd, 2003 to December 2nd, 2003.

2.1 WPPT forecasts

The data set consist of measurements of power production from Klim wind power plant and the predicted power from WPPT, based on three different weather forecasting systems. The installed power at Klim is 21000kW since 35 600kW V44 rotors are available and hence the predictions from WPPT range from 0 to 21000kW. Thus, this is the range available for the (absolute) prediction error as well.

The aim is to combine forecasts with the forecasts from WPPT as the explanatory variables, which are represented as

DWD: Predicted wind power based on meteorological forecasts from *Deutcher Wetterdienst*.

HIRLAM: Predicted wind power based on meteorological forecasts from *DMI-HIRLAM*.

MM5: Predicted wind power based on meteorological forecasts from *MM5*.

The three different forecasts are based on different meteorological data and since all predicting for the same event they are all quite correlated, approximately 0.85. The prediction horizon is also considered as a variable in the analysis. The issue of missing data can influence the horizon variable if the difference in number of observed values in each horizon is large. An investigation reveals that none of the forecasts have great difference in horizons such that it would influence the studies.

When the WPPT forecasts are combined in the following study, the combination is defined with a slash (/) between the constituent forecasts. In tables the combination is also noted by forecasts initials.

2.2 Meteorological forecasts

The data set consist of meteorological data from *DMI*'s meteorological forecasting system *DMI-HIRLAM*. The meteorological (MET) forecasts are only available at specific grid points over Denmark and to approximate the forecasts located at Klim, a bilinear interpolation between the four points around Klim is performed.

The aim is to generate a conditional parametric model of the weights in the combined forecast. The weights are wanted to be conditioning on some explanatory variables which

are found in a set of the meteorological forecasts. The meteorological variables available in the data set are the following:

ws10m: Wind speed forecast 10 meters above ground level (m/s).

wd10m: Wind direction forecast 10 meters above ground level (degrees).

rad: Radiation forecast (W/m^2)

fv: Friction velocity forecast (m/s).

ad: Air density forecast (g/m^2).

wsL··: Wind speed forecast in model level ··, the levels in the data set are 31, 38, 39 and 40 (m/s).

wdL··: Wind direction forecast in model level ··, the levels in the data set are 31, 38, 39 and 40 (degrees).

tkeL··: Turbulent kinetic energy forecast in model level ··, the levels in the data set are 31, 38, 39 and 40 ($1000m^2/s^2$).

The model levels are different pressure levels of the atmosphere. The numbers position the levels, such that with increasing number there is a decrease in height above ground. Not all MET variables are used in the analysis due to similarities. Correlation is strong between the wind speed variables, only **wsL31** show some difference from the dependency. Therefore two wind speed variables are considered, at 10m a.g.l. and at pressure level 31. The same is for the wind direction, and same levels are considered. The variables for turbulent kinetic energy are all alike and thus only one is applied, at level 38.

3 Modelling combined forecasts

Since the problem of combining was first presented many methods have developed ranging from a method of equal weights to more complex, e.g. state space and neural networks. Despite all these methods, an adoption of the linear model is most favorable when combining. Here, the linear model is generated where the weights are smooth but otherwise unknown functions.

In the study two performance measurements are used; the root mean square error and the coefficient of determination. The root mean square error (RMSE) is a error measure between a forecast and the actual values in the same units. Thus, it is a good measurement to visualize and is a good candidate when two or more forecasts are compared. The RMSE and other optional error measures are illustrated in Madsen et al. (2005b). The coefficient of determination (R^2) is a measure of total variability in the response accounted for by the model, that is R^2 is on the scale zero to one where improved fit indicates increase in proportion where one implies 100% fit. Discussion about R^2 and its adjustment can be seen in Montgomery and Runger (2002).

3.1 Linear models

Applied to wind power prediction the variable of interest, the predicted variable, is the actual wind energy production and at time t it is denoted as y_t. Let $\widehat{y}_{i,t}$ be the i-th individual forecast at time t, the prediction error between the production and the i-th competing prediction is

$$e_{i,t} = y_t - \widehat{y}_{i,t} \tag{1}$$

where $i = 1, \ldots, K$. A linear combination of the forecasts is then formulated as

$$\widehat{y}_{c,t} = w_{0,t} + \sum_{i=1}^{K} w_{i,t}\widehat{y}_{i,t} \tag{2}$$

where $w_{i,t}$ is the weight given to forecast i at time t. The term $w_{0,t}$ represents the constant in the linear model but inclusion of intercept reflects the bias of the individual forecasts. The combination model can be written as

$$y_t = \widehat{y}_{c,t} + e_{c,t} = w_{0,t} + \sum_{i=1}^{K} w_{i,t}\widehat{y}_{i,t} + e_{c,t}. \tag{3}$$

which gives the prediction error for the combination $e_{c,t} = y_t - \widehat{y}_{c,t}$. A vector notation for the combined forecasts is written as $\widehat{y}_{c,t} = \mathbf{w}_t^\top \widehat{\mathbf{y}}_t$ where $\widehat{\mathbf{y}}_t$ is a vector of K individual forecasts to be combined at time t and $\mathbf{w}_t$ is a vector of corresponding weights at time t. If included, constant is counted in the vector of weights and with one in the vector of forecasts. This means that by combining K number of forecasts the vector has $K + 1$ elements.

For larger prediction horizons than 1-step ahead, the h-step ahead prediction is calculated by

$$\widehat{y}_{t+h|t} = \mathbf{w}_t^\top \widehat{\mathbf{y}}_{t+h}. \tag{4}$$

In the following analysis the prediction is always at the same time and only one power prediction is available for every hour in the data. This implies that the updating in the recursive estimation takes place relative to each horizon, one step back means the last predicted value within the horizon. The notation h, indication of the prediction horizon, is thus omitted throughout the study.

3.1.1 Restriction and constant in the linear model

The parameters of interest are the weights which indicate the importance of an individual forecast in the combined forecast. The weighting contribute some fraction of information from each competing forecast to the combination and thus the weights are restricted to sum to unity, i.e.

$$\sum_{i=1}^{K} w_{i,t} = 1. \tag{5}$$

The modelling for the restriction is a straight forward method where the constraint is added into the combination model in (2) for the K-th weight and by subtracting with the

K-th forecast the combination model becomes

$$\widetilde{y}_t = w_{0,t} + \sum_{i=1}^{K-1} w_{i,t}\widetilde{y}_{i,t} + e_{c,t} \tag{6}$$

where $\widetilde{y}_{i,t} = \widehat{y}_{i,t} - \widehat{y}_{K,t}$ and $\widetilde{y}_t = \widehat{y}_t - \widehat{y}_{K,t}$. Neither the prediction error $e_{c,t}$ nor the constant $w_{0,t}$, if included, is affected by this modification. Not only the weights are consistently estimated, but also the restricted model in (6) give forecast errors for the linear combination model and corresponding intercept. What this restriction does not concern is the lower limit for the weights which is considered to be zero, inclusion of a constant in the model detects any strange behavior of the combination. The predictions from WPPT are for the power curve, so the diurnal variation has not been filtered from the forecasts. The inclusion of intercept can thus be interpreted as the diurnal variation for the forecasts.

In the following studies both constant and restriction in regression method is used. Counting for the restriction is equal to use the optimal procedure in combining but including the intercept takes out the bias of the individual forecasts.

3.1.2 Methods of combining

The methods applied in the following studies are the recursive least squares method and the adaptation of minimum variance-covariance. The performance of the simple average method is also applied for comparison. In Thordarson (2007) few more methods are briefly illustrated but not generated in the studies. The three methods applied in the following analysis are illustrated below:

Simple average The method is the simplest one and still today it appears in many situations to be the most consistent method of combining forecasts. Where K is the number of individual forecasts to be combined, the weights are all equal to

$$w_i = \frac{1}{K} \tag{7}$$

where the index i is a reference to individual forecast i in the combined forecast. Many applications have favored the simple average with the argument of it performing best or nearly best. It made Clemen (1989) ask why it works so well and under what conditions. A possible answer for the success of the method relies on unstable weights, which often result in unsystematic changes over time in the covariance matrix of the individual forecast errors (Holden and Peel, 1989).

Optimal The combination method is denoted as *optimal* when the individual weights are calculated to minimize the squared residuals of the combination, assumed that the individual forecasts are unbiased. This method is also called minimum variance-covariance method. The vector of combining weights, $\mathbf{w}$, is determined by the formula

$$\mathbf{w} = \frac{\mathbf{S}^{-1}\mathbf{u}}{\mathbf{u}^T\mathbf{S}^{-1}\mathbf{u}} \tag{8}$$

where $\mathbf{u}$ is the $n \times 1$ unit vector and $\mathbf{S}$ is the $n \times n$ covariance matrix of the forecast errors. More efficiency can be gained if the forecast errors were treated as independent. The weights are then depending on the variance of the individual forecast errors, which is observed as the diagonal terms of the covariance matrix, $\mathbf{V} = \text{diag}(\mathbf{S})$, and the weights are estimated by substitute $\mathbf{V}$ for $\mathbf{S}$ in (8).

The covariance matrix is time-varying which implies that the matrix $\mathbf{S}$ need to be estimated adaptively:

$$\widehat{\mathbf{S}}_t = (1 - \lambda)\mathbf{e}_t\mathbf{e}_t^\top + \lambda\widehat{\mathbf{S}}_{t-1}. \tag{9}$$

Some part of the data set is used to estimate the initial matrix $\widehat{\mathbf{S}}_0$ and to select the forgetting factor. An appropriate value for λ is between 0.95 and 0.999. The time-variation of the optimal method is quite well described in Sánchez (2006).

Regression In the most general form is the regression model of the combination written

$$y_t = g(\widehat{\mathbf{y}}_t, t; \mathbf{w}_t) + e_{c,t} \tag{10}$$

where $g(\widehat{\mathbf{y}}_t, t; \mathbf{w}_t)$ is a known mathematical function of the independent individual forecasts $\widehat{\mathbf{y}}_t = (\widehat{y}_{1,t}, \ldots, \widehat{y}_{K,t})^\top$, but the weights $\mathbf{w} = (w_{1,t}, \ldots, w_{K,t})^\top$ are unknown. The error $e_{c,t}$ is a random variables with $E[e_{c,t}] = 0$ and $V[e_{c,t}] = \sigma^2_{e_{c,t}}$. The general linear model is a special case of the regression model where the estimated response is a linear function on its parameters:

$$y_t = \mathbf{w}_t^\top \widehat{\mathbf{y}}_t + e_{c,t}. \tag{11}$$

To obtain weights for the most adequate model some loss function L is minimized. For the combined forecast to be competitive with the individual forecasts its loss function has to have equal or lower magnitude than the individual loss functions. The solution to the combination problem is a vector of weights, $\mathbf{w}_t = (w_{1,t}, \ldots, w_{K,t})^\top$, such that it minimizes the loss function $L(e_{c,t})$ in a way that $L(e_{c,t}) \leq \min_i \{L(e_{i,t})\}$. The loss function for the combination is the least squares function, $L(e_{c,t}) = E[(e_{c,t})^2]$. The LS method is then used to estimate the weights in the combination.

In the analysis both static and recursive approach for the weight estimation, is applied. In static estimation the whole data set is used to estimate the unknown time-invarying parameters. The design matrix is thus a matrix of the constituent forecasts including all values in the data set ($\widehat{\mathbf{y}}_t \to \widehat{\mathbf{y}}$ in (11)), and the measurements form a vector of observed values ($y_t \to \mathbf{y}$). The solution to the problem is an estimator of the weights that minimizes the sum of squared residuals,

$$\widehat{\mathbf{w}} = (\widehat{\mathbf{y}}^\top \widehat{\mathbf{y}})^{-1} \widehat{\mathbf{y}}^\top \mathbf{y}. \tag{12}$$

This is the solution to the ordinary least squares problem where no consideration is taken to residuals which may have larger variance or residuals which may be correlated. When it occur the problem is referred to as the weighted least squares problem where, in general, the estimator is

$$\widehat{\mathbf{w}} = (\widehat{\mathbf{y}}^\top \mathbf{\Sigma}^{-1} \widehat{\mathbf{y}})^{-1} \widehat{\mathbf{y}}^\top \mathbf{\Sigma}^{-1} \mathbf{y} \tag{13}$$

where $\widehat{\mathbf{y}}^\top \mathbf{\Sigma}^{-1} \widehat{\mathbf{y}}$ has full rank, i.e. at least one row in the matrix has k nonzero number of elements and therefore the matrix is invertable.

By estimating the weights adaptively the coefficients are allowed to be time-varying. For that the recursive least squares estimates are applied, which minimizes the weighted least squares estimator $\widehat{\mathbf{w}}_t = \arg\min S_t(\mathbf{w})$ where $S_t(\mathbf{w})$ denotes the quadratic loss function

$$S_t(\mathbf{w}) = \sum_{s=1}^{t} \beta(t,s) \left(y_s - \widehat{\mathbf{y}}_s^\top \mathbf{w} \right)^2. \tag{14}$$

The quantity $\beta(t,s)$ is the weight given to the s-th residual in the quadratic function at time t. Deviation between every two time consecutive β's give some value on λ relative to the time index. This value is called a forgetting factor and is considered to be constant throughout the following analysis. The loss function is then weighted exponentially as $\beta(t,s) = \lambda^{t-s}$ and the procedure reduces the importance of old data in the power of the forgetting factor. The RLS procedure with forgetting can be found to be

$$\widehat{\mathbf{w}}_t = \widehat{\mathbf{w}}_{t-1} + \mathbf{P}_t \widehat{\mathbf{y}}_t \left[y_t - \widehat{\mathbf{y}}_t^\top \widehat{\mathbf{w}}_{t-1} \right] \tag{15}$$

$$\mathbf{P}_t = \frac{1}{\lambda} \left[\mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1} \widehat{\mathbf{y}}_t \widehat{\mathbf{y}}_t^\top \mathbf{P}_{t-1}}{\lambda + \widehat{\mathbf{y}}_t^\top \mathbf{P}_{t-1} \widehat{\mathbf{y}}_t} \right]. \tag{16}$$

In order to obtain the recursive estimation it is important to provide an appropriate value on λ for the performance of the adaptive procedure, but an appropriate value is between 0.95 and 0.999. For further reading on the method of recursive least squares see Madsen (2001).

3.2 Nonlinear models

There is a class of models which are linear to some regressors, but the coefficients are assumed to be changing smoothly as an unknown function of some other variables. These kind of models are called *varying-coefficient models* (Hastie and Tibshirani, 1993), but when all coefficients depend on the same variable the model is referred to as *conditional parametric model*.

3.2.1 Locally weighted regression and conditional parametric models

Let y_i, for $i = 1,\ldots,n$, be the i-th measurement of the response and $\mathbf{x}_i$ be a vector of measurements of K explanatory variables at the same i-th moment. The model for local regression has the same basic structure as that for parametric regression in (10):

$$y_i = g(\mathbf{x}_i) + e_i, \tag{17}$$

where g is smooth function and e_i random error, i.i.d. and Gaussian. Assuming the function to be smooth allows points in certain neighborhood of $\mathbf{x}$ to estimate the response. This neighborhood is some fraction of the data closest to $\mathbf{x}$ where each point is weighted according to distance; increase in distance from $\mathbf{x}$ gives decrease in weight. The smoothing function is estimated by fitting a polynomial of the dependent variables to the response, $g(\mathbf{x}) = p(\mathbf{x},\mathbf{x})$. For each fitting point, parameters of the polynomial ($\boldsymbol{\theta}$) need to

be estimated, therefore locally-weighted least squares is considered:

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} w_i(\mathbf{x}) \left(y_i - p(\mathbf{x}_i, \mathbf{x})\right)^2. \tag{18}$$

The local least squares estimate of $g(\mathbf{x})$ is $\widehat{g}(\mathbf{x}) = \widehat{p}(\mathbf{x}, \mathbf{x})$. These local estimates are also called *local polynomial estimates*, but if the polynomial is of degree zero it is denoted as *local constant estimates*. The issue of locally weighted regression is the subject in Cleveland (1979) and Cleveland and Devlin (1988) where its properties are well explained.

The locally weighted regression requires a weight function and a specified neighborhood size. To allocate the weights $w_i(\mathbf{x})$ to the observations, a nowhere increasing weight function W is applied. There are several weight functions which can be used and some are listed in Table 1. In the case of spherical weight function the weight on observation i is determined by the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}$, i.e.

$$w_i(\mathbf{x}) = W\left(\frac{\|\mathbf{x}_i - \mathbf{x}\|}{h(\mathbf{x})}\right). \tag{19}$$

The positive scalar $h(\mathbf{x})$ is called the bandwidth. The bandwidth is an indicator for the neighborhood size. If constant for all value of $\mathbf{x}$ it is denoted as a *fixed bandwidth*. If $h(\mathbf{x})$ is chosen such that certain fraction (α) of the observations, $\mathbf{x}_i$, is within the bandwidth it is denoted as a *nearest neighbor bandwidth*. If $\mathbf{x}$ has dimension of more then one, scaling of the individual elements of $\mathbf{x}_i$ is considered before applying the method.

When using a conditional parametric model to formulate the response y_i, the explanatory variables are split in two groups. One group of variables $\mathbf{x}_i$ enter globally through coefficients depending on the other group of variables $\mathbf{u}_i$, i.e.

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}(\mathbf{u}_i) + e_i, \tag{20}$$

where $\boldsymbol{\theta}(\cdot)$ is a vector of coefficient functions to be estimated and e_i is the error term. The dimension of $\mathbf{x}_i$ can be quite large, but for practical purposes the dimension of $\mathbf{u}_i$ must be low. The functions $\boldsymbol{\theta}(\cdot)$ are estimated at a number of destinct points, fitting

Table 1: Some weight functions.

Name	Weight function
Box	$W(u) = \begin{cases} 1, & u \in [0,1) \\ 0, & u \in [1,\infty) \end{cases}$
Triangle	$W(u) = \begin{cases} 1-u, & u \in [0,1) \\ 0, & u \in [1,\infty) \end{cases}$
Tri-cube	$W(u) = \begin{cases} (1-u^3)^3, & u \in [0,1) \\ 0, & u \in [1,\infty) \end{cases}$
Gauss	$W(u) = \exp(-u^2/2)$

points, by approximating the functions using polynomials and fitting the resulting linear model locally to each of these points. Let $\mathbf{u}$ denote a particular fitting point, let $\theta_j(\cdot)$ be the j-th element of $\boldsymbol{\theta}(\cdot)$ and let $\mathbf{p}_{d(j)}(\mathbf{u})$ be a column vector of terms in the corresponding d-th order polynomial evaluated at $\mathbf{u}$. If for instance $\mathbf{u} = [u_1 \quad u_2]^\top$, then $\mathbf{p}_2(\mathbf{u}) = [1 \quad u_1 \quad u_2 \quad u_1^2 \quad u_1u_2 \quad u_2^2]^\top$. Also let $\mathbf{x}_i = [x_{1,i} \cdots x_{p,i}]^\top$. Then

$$\mathbf{z}_i^\top = \left[x_{1,i}\mathbf{p}_{d(1)}^\top(\mathbf{u}_i) \cdots x_{j,i}\mathbf{p}_{d(j)}^\top(\mathbf{u}_i) \cdots x_{p,i}\mathbf{p}_{d(p)}^\top(\mathbf{u}_i)\right] \tag{21}$$

$$\boldsymbol{\phi}_u^\top = \left[\boldsymbol{\phi}_{u,1}^\top \cdots \boldsymbol{\phi}_{u,j}^\top \cdots \boldsymbol{\phi}_{u,p}^\top\right], \tag{22}$$

where $\boldsymbol{\phi}_{u,j}$ is a column vector of local coefficients at $\mathbf{u}$ corresponding to $x_{j,i}\mathbf{p}_{d(j)}^\top(\mathbf{u}_i)$. The linear model

$$y_i = \mathbf{z}_i^\top \boldsymbol{\phi}_u + e_i \tag{23}$$

is then fitted locally to $\mathbf{u}$ using weighted least squares. The loss function which is minimized is

$$\widehat{\boldsymbol{\phi}}(\mathbf{u}) = \arg\min_{\boldsymbol{\phi}_u} \sum_{i=1}^{N} w_u(\mathbf{u}_i)\left(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}_u\right)^2, \tag{24}$$

for which a unique closed form solution exists, provided the matrix with rows $\mathbf{z}_i^\top$ corresponding to non-zero weights has full rank. The weights are the same as illustrated above in description on local estimates. The elements of $\boldsymbol{\theta}(\mathbf{u})$ are estimated by

$$\widehat{\theta}_j(\mathbf{u}) = \mathbf{p}_{d(j)}^\top(\mathbf{u})\widehat{\boldsymbol{\phi}}_j(\mathbf{u}) \tag{25}$$

where $j = 1, \ldots, p$ and $\widehat{\boldsymbol{\phi}}_j(\mathbf{u})$ is the weighted least squares estimates of $\boldsymbol{\phi}_{u,j}$. When $\mathbf{z}_j = 1$ for all j this method is identical to the locally weighted regression described above.

3.2.2 Adaptive estimation

If the estimates are defined locally to a fitting point $\mathbf{u}$, the adaptive estimates corresponding to this point can be expressed as

$$\widehat{\boldsymbol{\phi}}_t = \arg\min_{\boldsymbol{\phi}} \sum_{i=1}^{t} \lambda^{t-i} w_u(\mathbf{u}_i)\left(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}\right)^2 \tag{26}$$

where $w_u(\mathbf{u}_i)$ is a weight on observation i depending on the fitting point $\mathbf{u}$ and $\mathbf{u}_i$. The adaptive estimates in (26) can be found recursively as

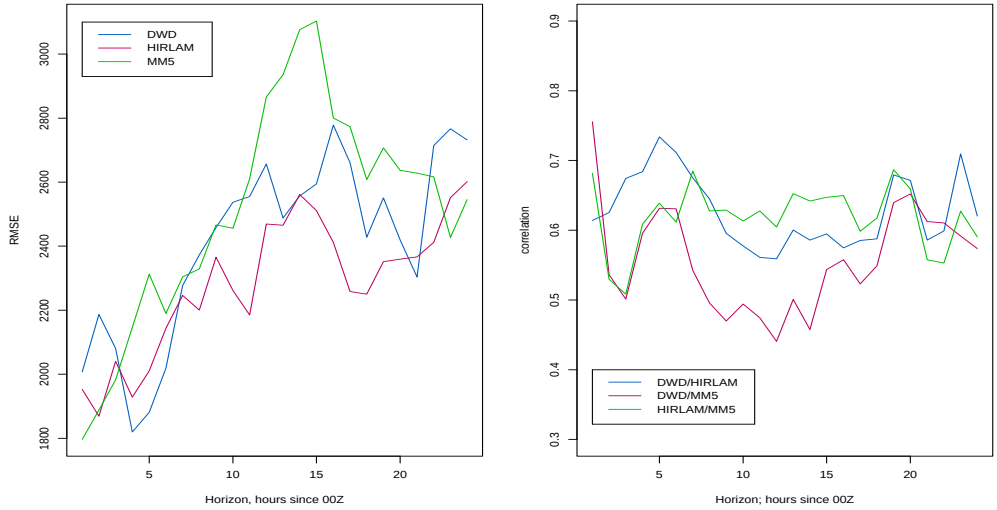
$$\widehat{\boldsymbol{\phi}}_t(\mathbf{u}) = \widehat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) + w_u(\mathbf{u}_t)\mathbf{R}_{u,t}^{-1}\mathbf{z}_t\left[y_t - \mathbf{z}_t^\top \widehat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})\right] \tag{27}$$

and

$$\mathbf{R}_{u,t} = \lambda \mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^\top. \tag{28}$$

Note that $\widehat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is a predictor of y_t locally with respect to $\mathbf{u}$ and for this reason it is used in (27). To predict y_t a predictor like $\widehat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is appropriate.

The method of adaptation for local estimation is not applied in this study. For more details on time-varying coefficient functions see Madsen and Holst (2000) or Nielsen et al. (2000).



(a) RMSE (b) Correlation

Figure 2: Individual forecasts.

4 Combining WPPT forecasts

The focus is on combining the WPPT forecasts, listed in Section 2.2 with the methods introduced in Section 3.1.

4.1 Individual forecasts

The individual forecasts are investigated to see if the conclusion drawn from the combination can be linked to the behavior of the individual predictions. The RMSE for the individual forecasts for all prediction horizons is depicted in Figure 2(a). RMSE is low for the shorter horizons, but increases when further away from the prediction time 00Z. From prediction horizon 7 the HIRLAM forecast is the best performing forecast. DWD forecast is almost as good as HIRLAM but varies more and for horizons 7 to 20 it is less accurate than HIRLAM. MM5 forecast is the least accurate prediction of all three and is bad representative for forecasting in horizon 11 to 15. It would be interesting to see which two competing forecasts is best to combine, especially between lead 11 to 20 since the difference in RMSE within forecasts appear to be the greatest in these horizons.

From Figure 2(a) it might be assumed that combining the two best performing forecasts result in the most adequate combination. This is not necessarily the case since RMSE is an overall measure for the accuracy in each horizon and does not concern the direction of each error term from the observations. However, the correlation is a good representative to compare inner structure of two processes, therefore it is interesting to see if the correlation of the prediction errors can give any knowledge about the best combination. The correlation between individual forecasts is shown in Figure 2(b) where all sets of

forecasts are quite correlated, around 0.6. It is though worth to notice that the HIRLAM forecast correlated with the other two, varies less than the correlation between DWD and MM5. The correlation from horizon 6 to 15 is lower between DWD and MM5 than for other combinations.

4.2 Offline combination

When the linear model for combining is considered the parameters are quantified as Figures 3 and 4 show for weights and an intercept, respectively. In the case of combining

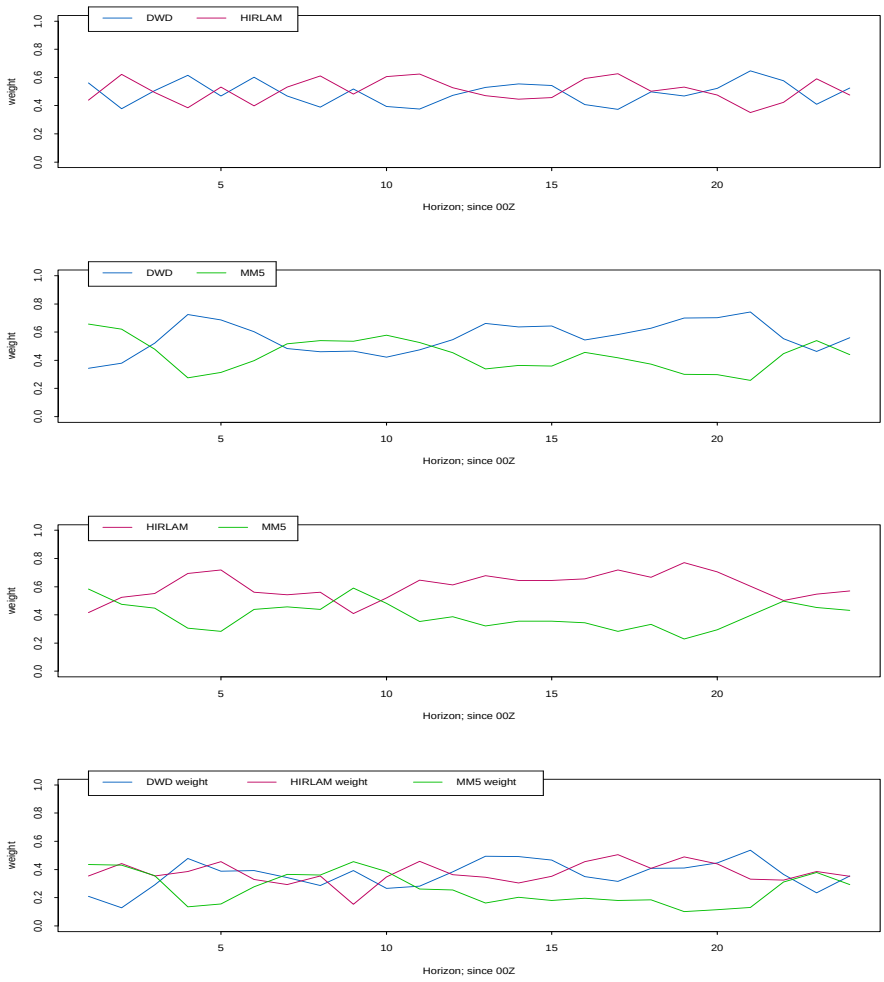


Figure 3: Size of the weights in a combined forecast. Each panel shows how the weights change with prediction horizons in a combination. The legends in the panels indicate what individual forecasts are combined.

DWD and HIRLAM the weights have similar behavior around 0.5 which is the mean weight. The constant term for the combination distinguish it from being comparable to the simple average method due to significance for all prediction horizons. Despite being

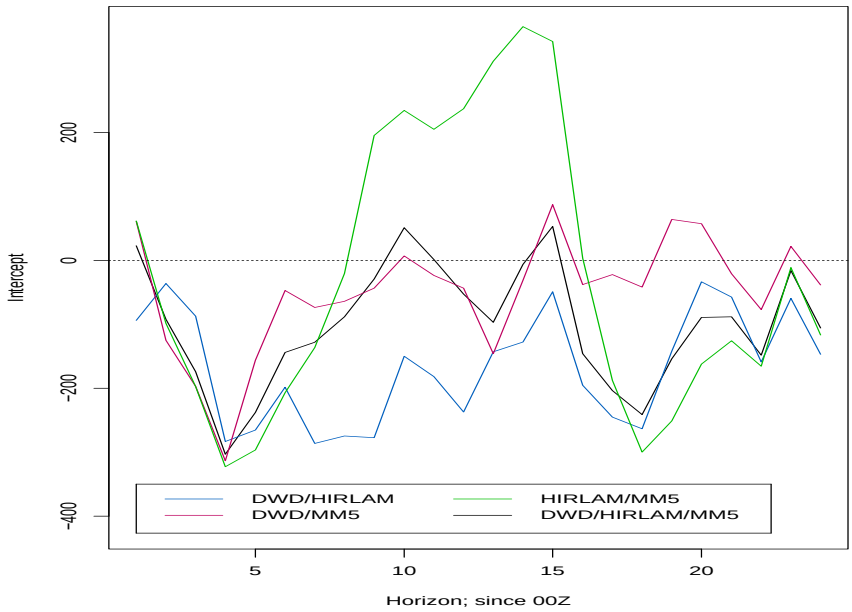


Figure 4: Estimated intercept in each horizon for the combined forecasts. The estimations can not be neglected.

the two best performing individual forecasts, the DWD/HIRLAM combination appears to have the lowest coefficient of determination as indicated in Table 2. For horizons listed, DWD/HIRLAM is the least fitted model for two forecast synthesis except for forecasting 18 hours ahead. The table also features the HIRLAM/MM5 forecast to be the best performing aggregation for the first horizons, and DWD/MM5 outperforming the others from prediction horizon 6. Adding the third forecast into the combining increases the coefficient of determination for all horizons, which indicates that a more precise model is gained by including additional forecasts.

Figure 5 shows RMSE for all offline combinations in all 24 prediction horizons. It confirms the best performance of the HIRLAM/MM5 forecasts in the shortest horizons till DWD/MM5 become the best performing prediction. However, what Figure 5 shows is

Table 2: An in-sample coefficient of determination (R^2) for the whole data set. The estimated weights from the restricted model are used in the linear model to determine R^2.

Combination	Prediction horizon [hours]						
	1	2	3	6	12	18	24
D/H	0.770	0.796	0.791	0.781	0.817	0.724	0.681
D/M	0.807	0.827	0.829	0.821	0.847	0.758	0.702
H/M	0.818	0.848	0.834	0.818	0.838	0.689	0.701
D/H/M	0.822	0.850	0.844	0.833	0.862	0.758	0.727

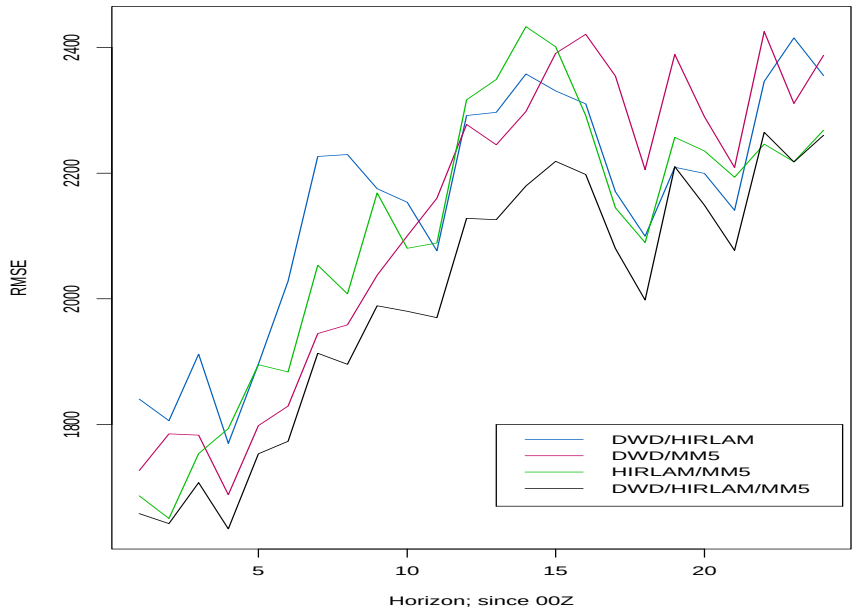


Figure 5: In-sample RMSE for combined forecasts over the prediction horizons in an offline estimation.

that for horizon 19 to 21, DWD/HIRLAM outperforms the other two. From lead 10 the two combination including MM5 are performing similarly. For the entire prediction horizon, the combination of all three competing forecasts outperform the composite of any two predictions. This is not surprising since information from all three are gathered to improve the accuracy. It is however noticed from Figure 5 that for the first few horizons and the few last, the performance of DWD/HIRLAM/MM5 is only slightly better than the best performing synthesis of two forecasts.

4.3 Online combination

The choice of an appropriate forgetting factor is a key feature of adaptation since it has a substantial effect on the efficiency of the predictions. The normal procedure for selecting the forgetting factor is to use the first part of the data set for the choice and then use the found λ for the whole data set, but missing values in a data set can influence the selection. Therefore is it more sufficient to use the longest period of non-missing values in the data set for the evaluation. Combined DWD and HIRLAM has non-missing values up to 150 days ahead, but with MM5 included in combination has only 56 days without missing data. Thus, evaluating λ for more the 56 days might be influenced by the missing values. The search for the forgetting factor concluded that the minimum distance between the prediction and the observed values appears when past 50 daily observations are used to estimate the weights. These 50 days give forgetting factor of $\lambda = 0.98$ and is the objective for all methods and possible combinations.

Figure 6 shows the first, intermediate and last three time-varying weights for each forecast weight. The DWD weights are displayed with various starting weights, but with time the coefficients become stable where all horizons have similar weights. It is only when DWD is combined with HIRLAM that the weights differ where DWD forecast have more effects on shorter horizons. All possible combinations with HIRLAM show the same structure where HIRLAM has small influence within corresponding combined forecast for shorter horizons, but the influence increases for larger prediction horizons.

With three individual forecasts there are equally many options of combining two forecasts. Figure 7 shows the results from combining two forecast with RLS method, compared with the performance of the individual forecasts. It illustrates that great reduction in distance from actual observations is accomplished by combining. All combinations are more accurate than any of the competing predictions. The figure also shows that the two best performing individual forecasts give the least performing aggregation. It is only for prediction horizon 15 and 18 to 21 that the DWD/HIRLAM is the most beneficial synthesis. The DWD/MM5 forecast is the best combination for the first half of the prediction horizons and in the latter half, combinations including HIRLAM are more precise.

The recursive estimation was also performed with the optimal procedure. By including the intercept in the regression the issue of bias in the individual forecasts is partly omitted[1] in the combination. If the intercept appears to be close to zero it can be neglected and the optimal method would perform as well as the adaptive regression. What Figure 8 illustrates is a significant difference in the accuracy for the two adaptive methods in favor of regression for all prediction horizons. The inclusion of constant term in the combination model can not be ignored in any of the three possible synthesis.

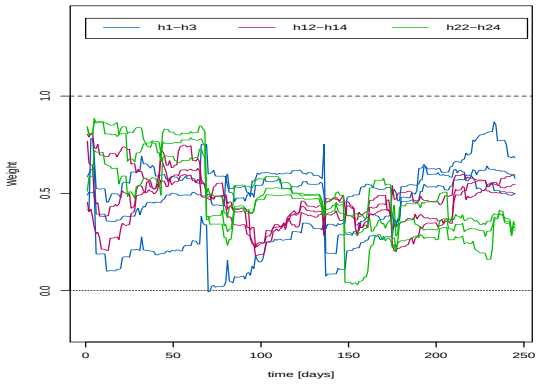
In Figure 9 the combined forecast with three individual forecasts is displayed for both RLS and optimal method along with combination of two forecasts with RLS procedure. Additional information from the third forecast reduces RMSE even further. The improvement is most in prediction horizons 12 to 16 which are the horizons where most deviation in accuracy of individual forecasts appears. The same difference as before is visible between the optimal method and recursive least squares method when the third prediction is augmented to the composition.

By comparing Figures 5 and 9 the importance of estimating the weights adaptively is visualized. Great improvement in accuracy is achieved along with the ability of detecting strange behavior in the time-varying weights.

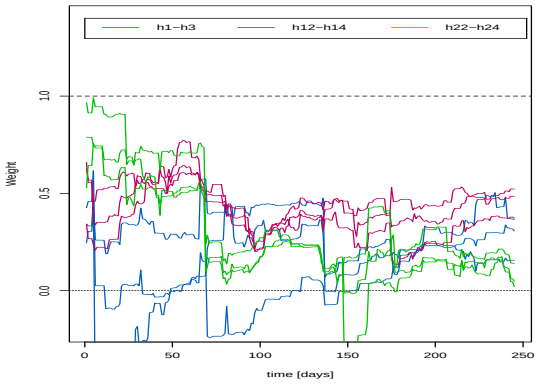
Table 3 shows a coefficient of determination for selected horizons for three different methods. It illustrates the supremacy of the recursive least squares method over the optimal and Simple Average method (SA). For all horizons depicted in the table, RLS method outperforms other methods. It also confirms the results from Table 2 about the individual forecasts, the least performing combination includes the forecasts with lowest RMSE individually (DWD/HIRLAM).

The correlation between every two competing forecast errors appears to give some idea about the combination. If two power forecasts are highly correlated the distance from the actual power production to these forecasts is the same both in magnitude and direction.

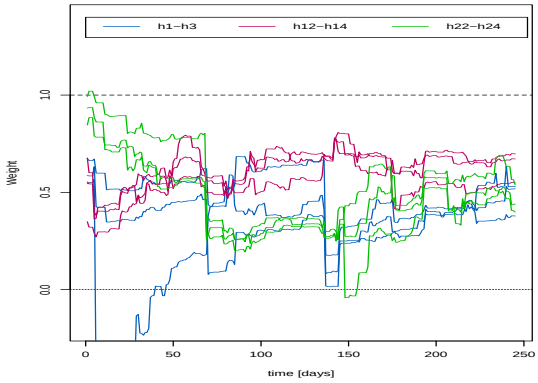
[1] de Menezes et al. (2000) claims that the constant will only debias for location bias, but not scale bias.



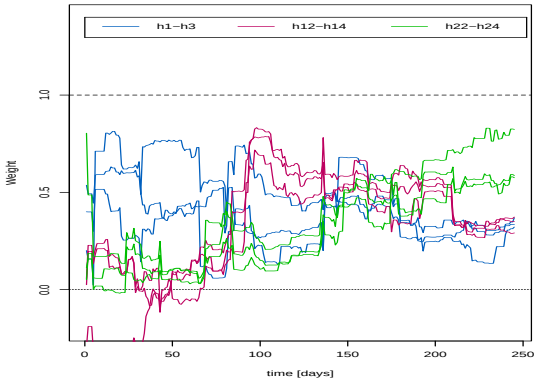
(a) DWD weights in D/H



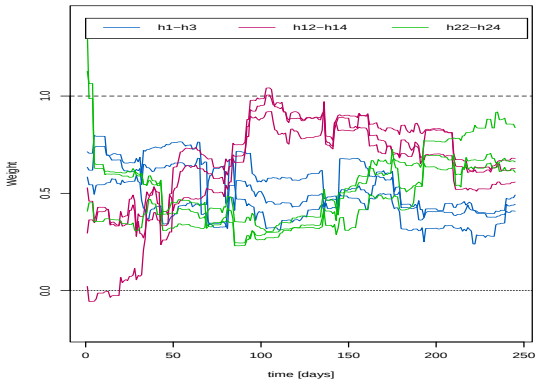
(b) DWD weights in D/H/M



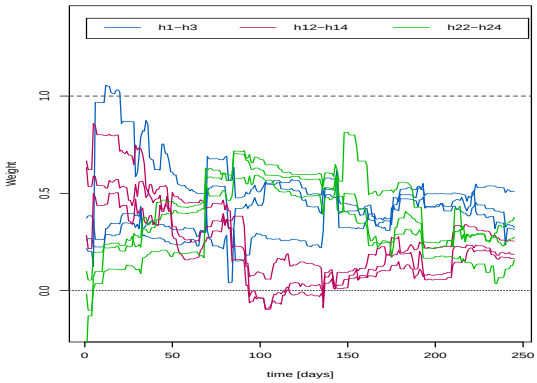
(c) DWD weights in D/M



(d) HIRLAM weights in D/H/M



(e) HIRLAM weights in H/M



(f) MM5 weights in D/H/M

Figure 6: First, intermediate and last time-varying weights for 4 combined forecasts. The second weight for a combination of two forecasts is a mirror of the first one through 0.5.

To be able to improve accuracy a forecast which appear on the opposite direction of the observed production is needed to approach the observations. Forecast errors on either

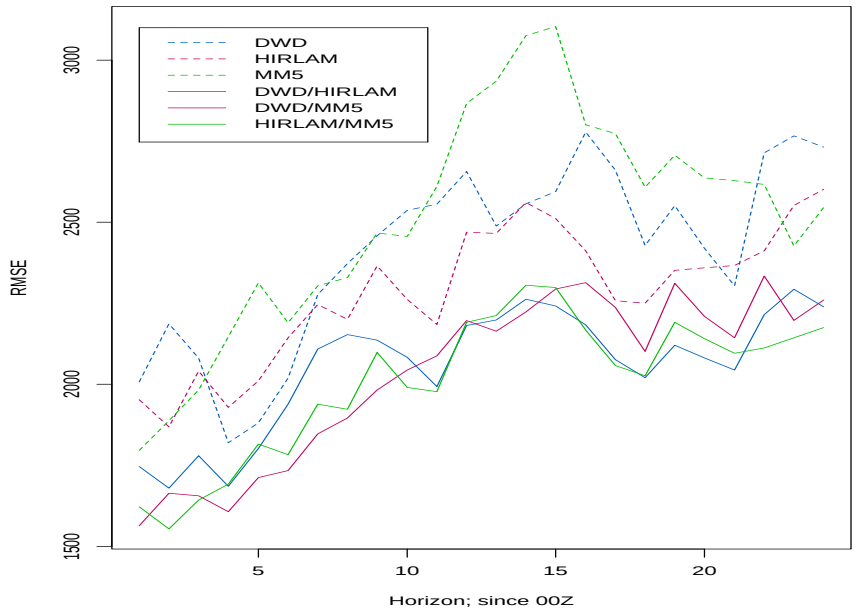


Figure 7: RMSE for combination of two forecasts with RLS method, compared to performance of the individual forecasts.

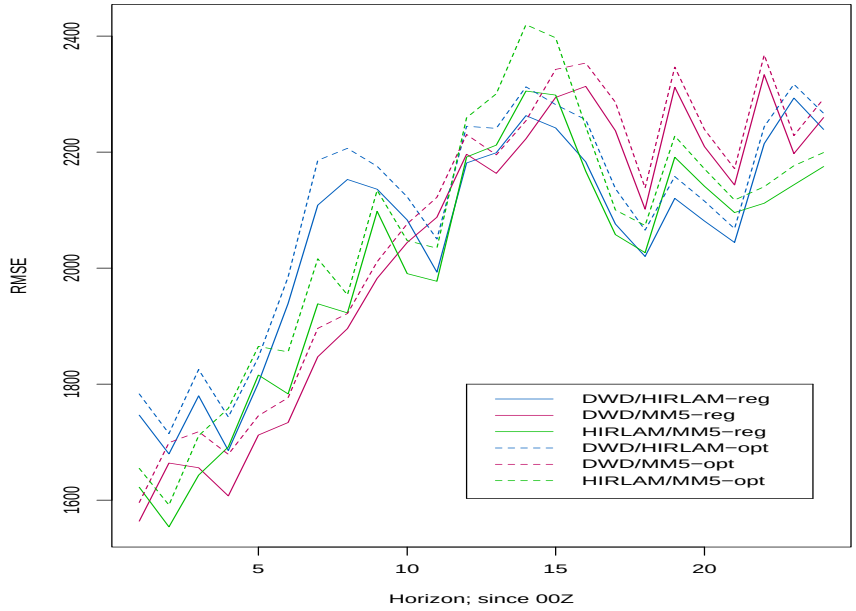


Figure 8: Performance comparison between RLS and OPT method when two forecasts are combined.

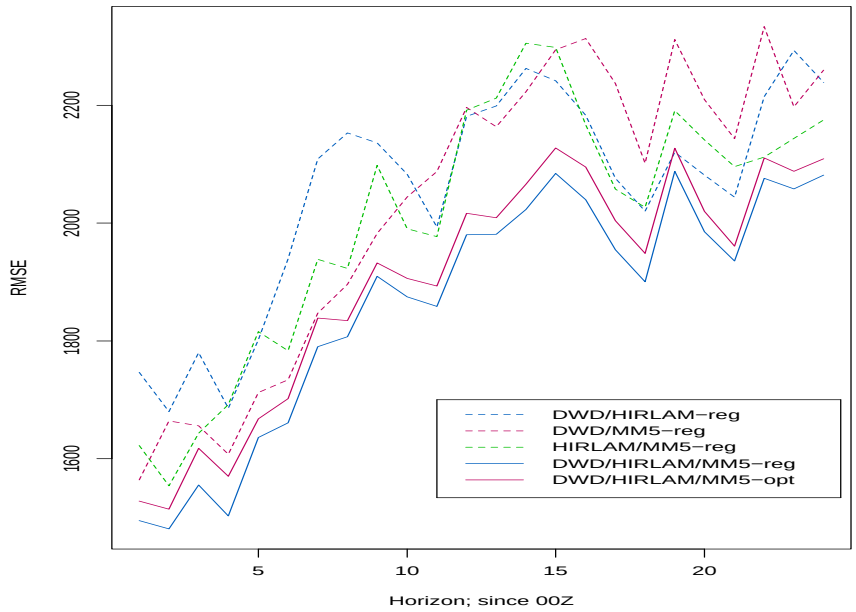


Figure 9: Two methods of combining three wind power forecasts compared with RLS method for two predictions combined.

direction of the power production would reduce the correlation. The correlation between the forecast errors in Figure 2(b) shows the DWD/MM5 having the smallest correlation over the intermediate horizons. The combination of these two forecasts gives the best

Table 3: Coefficient of determination (R^2) for combining forecasts with 3 alternative methods. The results are shown for selected prediction horizons between 1 hour and 24 hours.

Combination		Prediction horizon [hours]						
		1	2	3	6	12	18	24
RLS	D/H	0.803	0.815	0.810	0.815	0.838	0.812	0.694
	D/M	0.861	0.840	0.854	0.869	0.855	0.817	0.726
	H/M	0.850	0.860	0.856	0.862	0.855	0.829	0.746
	D/H/M	0.873	0.873	0.871	0.880	0.882	0.850	0.768
OPT	D/H	0.795	0.807	0.800	0.807	0.828	0.805	0.687
	D/M	0.855	0.833	0.843	0.863	0.850	0.810	0.718
	H/M	0.845	0.854	0.844	0.850	0.847	0.822	0.740
	D/H/M	0.867	0.868	0.861	0.874	0.877	0.843	0.761
SA	D/H	0.780	0.782	0.780	0.793	0.819	0.793	0.662
	D/M	0.827	0.809	0.829	0.853	0.843	0.797	0.698
	H/M	0.837	0.841	0.835	0.845	0.833	0.810	0.727
	D/H/M	0.842	0.836	0.842	0.863	0.862	0.829	0.729

combined forecast from two constituent predictions.

5 Fitting weights with local regression

The linear model for combining forecasts is a model which can be fitted with local regression. The weights from the regression can be extanded to get improvement in the combination by fitting the parameters by not only considering the past data, but the "future" as well by local regression. This is not an adaptive procedure, but can be considered as illustrated in Section 3.2.2.

In local regression, each fitting point on smoothed regression surface uses some fraction of the data set to estimate the fit. The fraction has to be chosen as large as possible to minimize the variability in the smoothing without twisting the pattern in the data. This fraction is exploited to the local regression by the bandwidth selected.

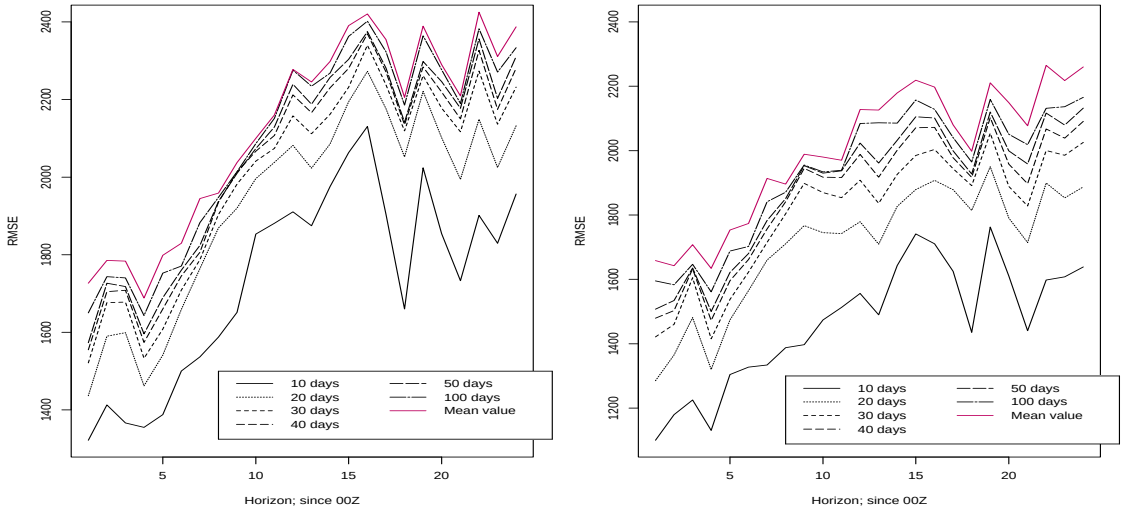
5.1 Bandwidth selection

The forgetting factor chosen in the RLS estimation in Section 4 represents the past days used for present estimation. The bandwidth in local regression is quite similar factor. It indicates how many data points are used in estimation, but the bandwidth considers data points in both directions from the fitting point. For the fitting points the bandwidth is considered to be fixed since the data is equally distributed over time.

The selection of bandwidth has a tradeoff between variance and bias. For low values of bandwidth the span for the estimation is short and the actual observed value is approached. This will decrease the bias in the estimation but narrowing close to the actual value will increase the variance. Extanding the bandwidth would reduce the variance as the bandwidth increases until it spans the whole data set. The smoothed value is then the mean of the observations which are fitted locally. This phenomena is illustrated in Figure 10. For each horizon RMSE increases with extension in the bandwidth. The red line in the panels is the mean value and is the upper limit for the bandwidth. This value is the one estimated for the offline estimation in Section 4.2. The panels also show how rapidly the RMSE increases for lower bandwidths. Around $h = 40$ (days) the rise almost vanish and the addition of single day to the bandwidth, gives little extension to the performance of the fit.

5.2 Comparison with RLS

The forgetting factor in RLS estimation was approximated 50 days and by estimating the weights over equal number of days, 25 days are selected in each direction for the bandwidth. In Figure 11(a) the local fit with bandwidth of 25 days in either direction is compared with RLS method. Quite definite improvement in performance is identified for all possible combinations, specially for large prediction horizons. In the smaller horizons the local fit and the RLS method perform very similarly. In Figure 12, four horizons



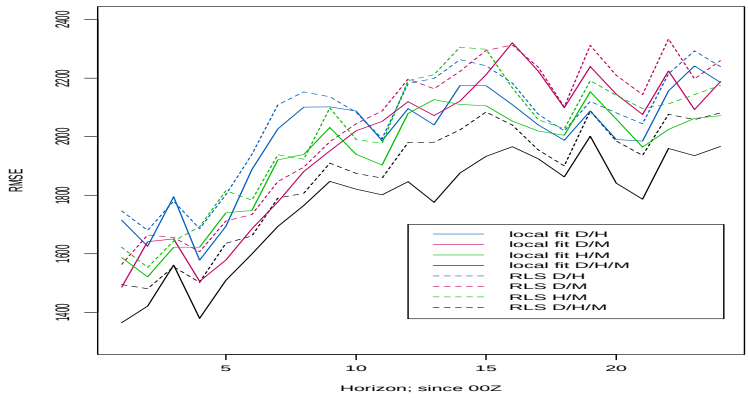
(a) DWD/MM5 (b) DWD/HIRLAM/MM5

Figure 10: For increasing bandwidth, RMSE increases. The red line indicates the mean value of the observed values in the data set, which is the upper limit for the local fit.

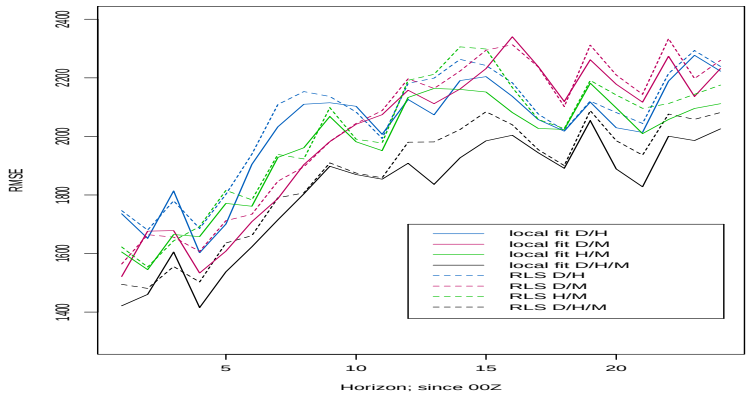
from the DWD/MM5 combination are displayed to illustrate how the local fit proceeds compared to the time-varying weights from RLS estimation. By using data close to a fitting point instead of only considering the previous information, a phase error in the recursively estimated weights is detected. The phase error can be seen by comparing the weight estimation from RLS and the locally fitted weights. For the weights, phase error is apperant where changes in the parameters are detected up to 25 days ahead. The sensitivity of abrupt changes is also noticed where expanding bandwidth reduces the changes in every step of the local fit, but accuracy of the estimation is sacrificed. This is due to the tradeoff between variance and bias for the local regression as illustrated above.

The increase in the bandwidth reduces the amplitude of the local fit until, eventually, it forms a straight line through the mean estimated weight. The comparison between the RLS estimates and the local fit with bandwidth 2×40 appears to be quite alike where the phase error aparts the estimated weights. Using local regression with bandwidth of 2×40 the recursive estimation is not outperformed as Figure 11(c) indicates. The performance appears to be similar for the methods within all combinations. The local regression should outperform the RLS estimation which indicates that the bandwidth has to be reduced. For $h = 2 \times 30$ the local fit is improved such that it is very similar to the performance of the recursive method over the 24 hour prediction horizon. This is depicted in Figure 11(b).

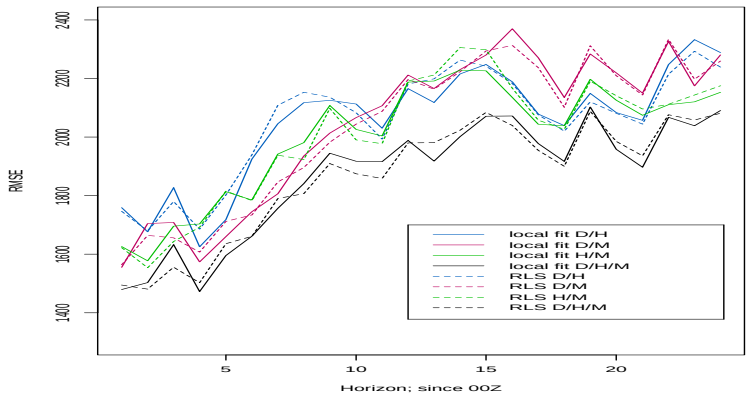
Table 4 shows the coefficient of determination for the local fit with bandwidth 60 days and 80 days, compared with the RLS fit from Table 3. Performance of local fit with bandwidth of 80 days is worse than RLS performance, but by reducing the bandwidth about 20 days the locally fitted performance improves RLS or is equivalent. By reducing the bandwidth



(a) $h = 2 \times 25$



(b) $h = 2 \times 30$



(c) $h = 2 \times 40$

Figure 11: RMSE of local regression with different bandwidths compared to RLS estimation.

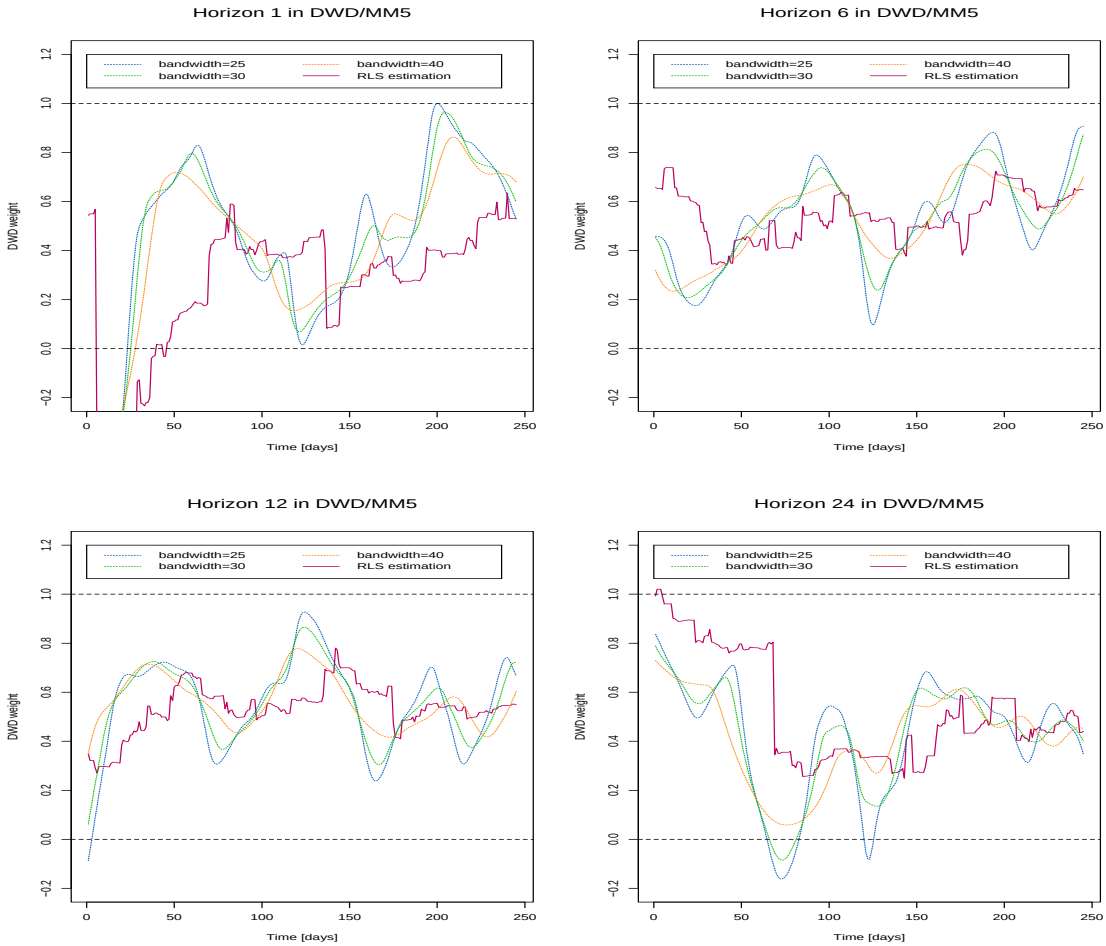


Figure 12: Few examples to illustrate how the bandwidth changes compared with the weights from RLS.

to 2×25 the local fit outperforms the RLS performance for all combinations. The only exception is the performance in the three hour horizon. Thus, the bandwidth for the local regression is fixed and selected close to 50 days to outperform the corresponding recursive least squares method.

6 Weight estimation using MET forecasts

The objective is to estimate weights in combined forecasts with informations from the MET forecasts. In previous sections the weights have been estimated by various methods including local regression. In the following the local regression will be evolved where the weights depend on one or more of the MET forecasts. The locally fitted weights are then included in the combination model, illustrated in (2), which will give a conditional parametric model of the combined forecast.

The following analysis focus on combining two forecasts with restriction. This has the simple approach of the forecast weights being linear dependent and the pattern which appears for one weight, is the same for the other.

6.1 Dependency between weights and MET forecasts

In Section 5 the conclusion is that using bandwidth close to 50 days gives an appropriate weights in the local fit. To find any relationship between the weights and the MET forecasts, a linear model is generated where the weights from the local regression are depending on one or more of the MET variables. The linear model is the general linear model explained in (11) with different notation, or

$$\mathbf{w} = \boldsymbol{\beta}^\top \mathbf{X} + \mathbf{e} \tag{29}$$

where $\mathbf{w}$ is a vector of weights from the local regression, $\mathbf{X}$ is a matrix with all explanatory variables, the MET forecasts in this presentation, and $\boldsymbol{\beta}$ are the coefficients to be estimated. The term $\mathbf{e}$ is a vector of residuals with mean zero and $\sigma_e^2 = 1$. By inspecting the scatterplots in Figure 13 it is difficult to see trends between the DWD weights and the MET forecasts. The red line in the plots is locally weighted regression between corresponding MET variable and the DWD weights. The weights seem to have some correlation with air density (**ad**) and turbulent kinetic energy (**tke**).

The disadvantage of using scatterplots to inspect dependent variables conditioned on explanatory variables, is that it only shows coherency with one variable. Relation of response variable with two explanatory variables can be demonstrated by *coplots* which is well illustrated by Cleveland (1994) in relation with conditionally parametric fits. One

Table 4: Coefficient of determination (R^2) for combining forecasts. Comparison between two locally fitted procedures and RLS performances. The results are shown for selected prediction horizons between 1 hour and 24 hours.

Combination		Prediction horizon [hours]						
		1	2	3	6	12	18	24
$h = 2 \times 30$	D/H	0.805	0.824	0.801	0.820	0.847	0.813	0.704
	D/M	0.885	0.877	0.863	0.886	0.890	0.852	0.780
	H/M	0.854	0.862	0.853	0.865	0.863	0.830	0.761
	D/H/M	0.885	0.877	0.863	0.886	0.890	0.852	0.780
$h = 2 \times 40$	D/H	0.800	0.819	0.798	0.816	0.840	0.809	0.687
	D/M	0.876	0.870	0.859	0.880	0.881	0.847	0.765
	H/M	0.850	0.856	0.848	0.862	0.856	0.827	0.751
	D/H/M	0.876	0.870	0.859	0.880	0.881	0.847	0.765
RLS	D/H	0.803	0.815	0.810	0.815	0.838	0.812	0.694
	D/M	0.861	0.840	0.854	0.869	0.855	0.817	0.726
	H/M	0.850	0.860	0.856	0.862	0.855	0.830	0.746
	D/H/M	0.873	0.873	0.871	0.880	0.882	0.850	0.768

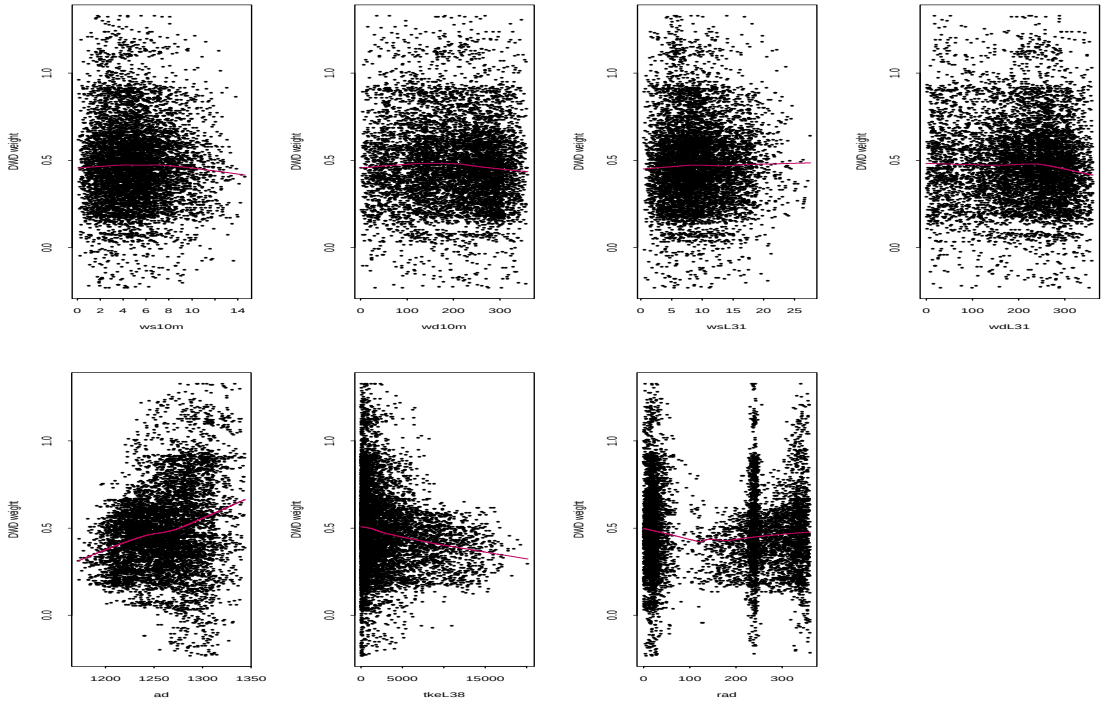
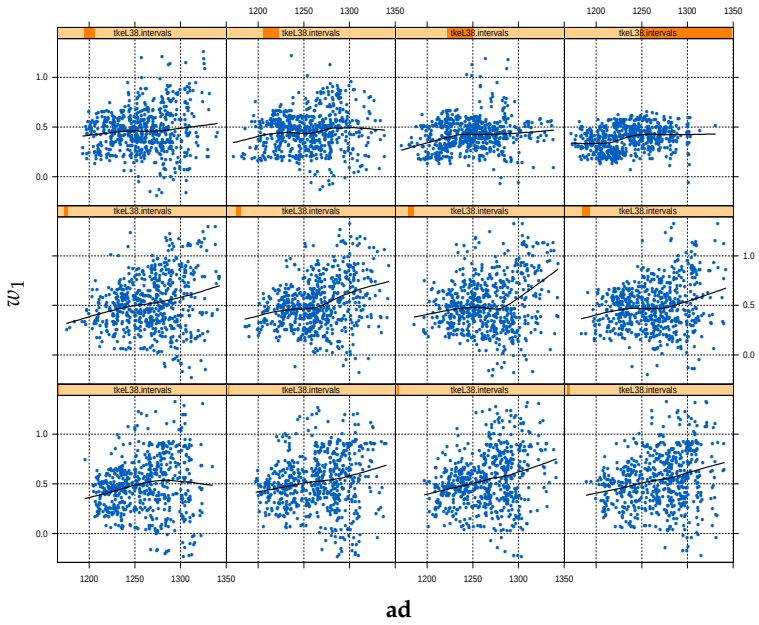


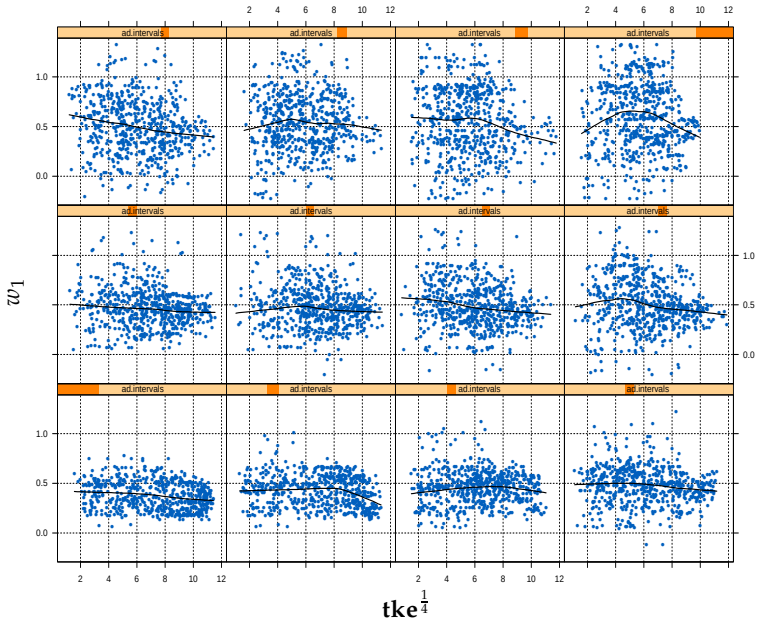
Figure 13: Scatterplot of the DWD weight in DWD/HIRLAM in relation with the MET forecasts.

explanatory variable is partitioned in several categories and the response variable is smoothed w.r.t. the other explanatory variable within the partitioning. Figure 14 shows coplots with the weights as the response and air density (**ad**) and turbulent kinetic energy (**tke**) as predictors. In the panels equally many points are used for the local regression. In Figure 14(a) **tke** is partitioned. It shows how the slope of the local fit between weights and **ad** decreases with higher values in **tke** as well as the variance of the weights. For low values in **tke** (bottom row of panels) little or no changing in slope occur but that is due to the density of the variable which is very dense for low values. When the turbulent kinetic energy increases the slope decreases to zero after the intermediate panels show some shifting in slope of the local fit around **ad**=1270. The distribution of the points in each panel also shows how the response spreads with increasing **ad**, but the distribution is reduced with increace in **tke**. By partitioning now the air density and fit the weights to the fourth root of **tke** (Figure 14(b)), the local regression appears to be constant for almost all panels. By inspecting the fit closer it can be seen that the fitted line shifts upwards with increase in air density. It is also noticed from the top row of panels that the fit generates a negative slope til it forms a concave curvature. The data points in the panels verify the distribution of the weights with increase in air density, while the variance of w_1 is constant with changes in **tke** within the panels. There seems to be some connection between these two MET forecasts and the DWD weights. The **ad** and **tke** forecasts are now applied as predictors for the weights in the conditional parametric model

$$\widehat{y}_c = w_0(\mathbf{ad}, \mathbf{tke}) + w_1(\mathbf{ad}, \mathbf{tke})\widehat{y}_1 + w_2(\mathbf{ad}, \mathbf{tke})\widehat{y}_2 \tag{30}$$



(a) Scatterplot of weights and **ad** where **tke** is partitioned.



(b) Scatterplot of weights and $\mathbf{tke}^{\frac{1}{4}}$ where **ad** is partitioned.

Figure 14: Coplots where DWD weights are depending on air density and turbulent kinetic energy.

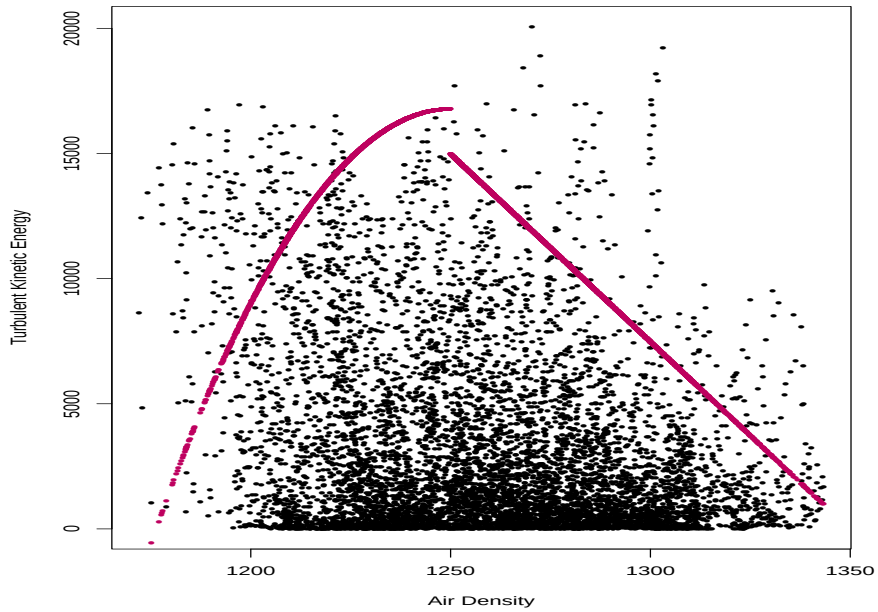


Figure 15: Scatterplot of **ad** *and* **tke** *variables. These variables make the basis for the surface of the weight estimation. The red lines define the convex hull.*

where the constraint for forecast coefficients, equation (5), is considered. The scatterplot of **ad** and **tke** makes a basis for the weight estimation in (30) but by observing the **ad**/**tke**-plane in Figure 15 it is seen that the MET data does not cover the whole plane. The **tke** forecast is more dense at low values and then diffuses when it increases, while the **ad** forecast is closer to be normally distributed. This implies that for high values of **tke** and either high or low values of **ad**, no or few observed values exist. The convex hull[2] is thus defined as the area inside the red lines in Figure 15, that is the area where the weights have some valid estimation on the basic plane.

Figure 16 shows the time series for the two MET forecasts which are of interest, for every hour from February to beginning of December. Air density is known to follow the behavior of air temperature and from the plot it is quite patent, the air density is high through the cold months of the year but decreases during the summer period. However, the turbulent kinetic energy is not correlated with other weather phenomenae of the atmosphere. It varies through the entire period, less though in both tails of the data set which indicates reduced variation over the winter period. Turbulent kinetic energy is a variable used to study turbulence and its evolution in boundary layers of air in the atmosphere. When the layers become stable the **tke** is suppressed. The time series plot implies that layers of air in the atmosphere are more stable over the winter months.

[2]Convex hull for a set of points X in a real vector space V is the minimal convex set containing X. (*www.wikipedia.org*)

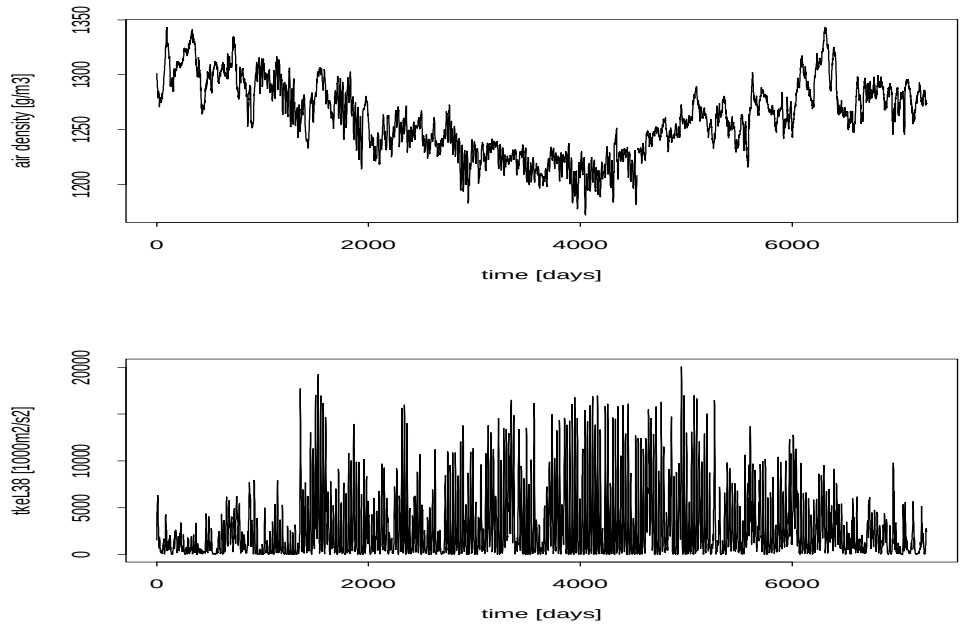


Figure 16: Time series plots for the air density and the turbulent kinetic energy at level 38

6.2 Using MET variables in local regression

Weights for all three possible combinations of two individual forecasts are examined on the **ad**/**tke**-plane. Only one weight from each combination is displayed since the weights are restricted to sum to one. As illustrated in Section 3.1.1, in the case of restriction on the parameters, the model used to estimate the weights can be rewritten as

$$\widehat{y}_c - \widehat{y}_2 = w_0(\mathbf{ad}, \mathbf{tke}) + w_1(\mathbf{ad}, \mathbf{tke})\,(\widehat{y}_1 - \widehat{y}_2) \tag{31}$$

when combining two individual forecasts and with the weights as a function of the two MET forecasts.

Locally-weighted linear model is considered for w_1, but for the constant term w_0 a local constant is approximated. The local models are

$$w_0(\mathbf{ad}, \mathbf{tke}) = w_0 \tag{32}$$

$$w_1(\mathbf{ad}, \mathbf{tke}) = w_{10} + w_{11} \cdot \mathbf{ad} + w_{12} \cdot \mathbf{tke}. \tag{33}$$

By substituting (32) and (33) into (31) a modified linear model for combination is

$$\begin{aligned} \widehat{y}_c - \widehat{y}_2 = w_0 &+ w_{10}\,(\widehat{y}_1 - \widehat{y}_2) \\ &+ w_{11} \cdot \mathbf{ad}\,(\widehat{y}_1 - \widehat{y}_2) \\ &+ w_{12} \cdot \mathbf{tke}\,(\widehat{y}_1 - \widehat{y}_2)\,. \end{aligned} \tag{34}$$

where the explanatory variables are now products of prior variables and the MET forecasts. The modified model in (34) includes new explanatory variables which are the

product terms. Instead of estimating two parameters, four are evaluated in the modified formulation. Combining wind power forecasts as in (34) indicates that the weights are linearly depending on the two MET forecasts. But the weights are unknown functions of the MET data and by smoothing the weights over the **ad**/**tke**-plane, the contourplots in Figure 17 are obtained. The weights on DWD when combined with HIRLAM or MM5, appear to have similar surfaces. For low values of **tke** the DWD weights become more effective with increase in **ad**, but for higher values on the turbulent kinetic energy the weights are reduced with progressing air density. The behavior of the HIRLAM weight when combined with MM5 is more challenging to interpret. There is some fluctuation for the low values of **tke**, but with increasing turbulent kinetic energy the surface gets more smooth.

The surfaces for the intercepts are quite similar where low **tke** implies high negative value for the intercept, but with increase in **tke** the intercepts increase as well. The intercepts depending on **ad** show some kind of bell-shaped structure where the high and low values on air density imply low value on intercept, but around mean **ad** the intercept is at its maximum.

6.2.1 Extension to conditional parametric model

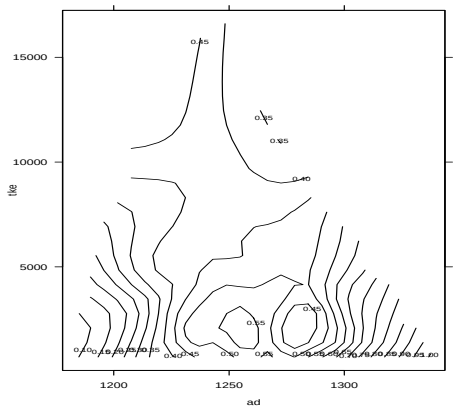
For the DWD weight in DWD/HIRLAM combination it can be concluded that there is linear relationship between weight and air density where both intercept and slope are functions of the turbulent kinetic energy:

$$w_1(\mathbf{ad}, \mathbf{tke}) = v_{10}(\mathbf{tke}) + v_{11}(\mathbf{tke}) \cdot \mathbf{ad}, \tag{35}$$

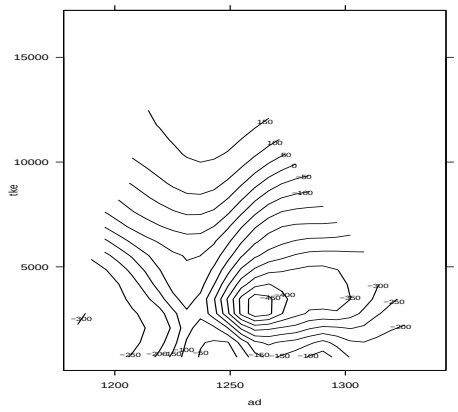
where v_{10} and v_{11} are the intercept and the slope respectively. This can be detected from the contourplot in Figure 17(a) or by the surface plot in Figure 18(a) where the approximated linearity can be visualized. The plots show that for increase in **tke** the intercept increases, but the slope decreases and become negative for **tke** higher than 8000. Observing the intercept $w_0(\mathbf{ad}, \mathbf{tke})$ shows a wavelike behavior for low values of **tke** around the mean value of **ad**. This might be challenging to interpret in a model presenting the intercept. The influence of the variance of the intercept can be estimated by comparing the terms of the conditional parametric model (CPM) in (31), e.g. the intercept and the product of the forecast weight and the forecast. Table 5 shows the covariance matrix of these terms and it indicates that the product between the weight and the predictor is about 8 times greater than the variance of the intercept. The variations for low **tke** on the surface of the intercept are therefore omitted.

Table 5: Covariance matrix for terms in (31) where $\widetilde{\widehat{y}}_1 = \widehat{y}_1 - \widehat{y}_2$.

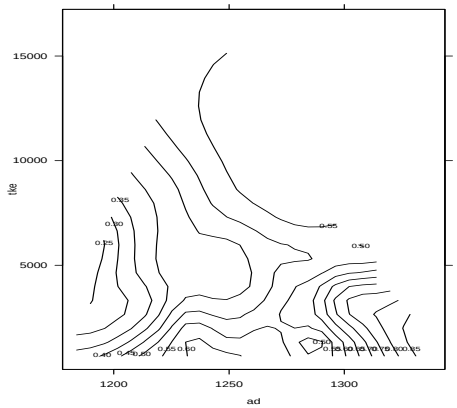
Variance	$w_0(\mathbf{ad}, \mathbf{tke})$	$w_1(\mathbf{ad}, \mathbf{tke})\widetilde{\widehat{y}}_1$
$w_0(\mathbf{ad}, \mathbf{tke})$	17369.5	24039
$w_1(\mathbf{ad}, \mathbf{tke})\widetilde{\widehat{y}}_1$	24039	1129552



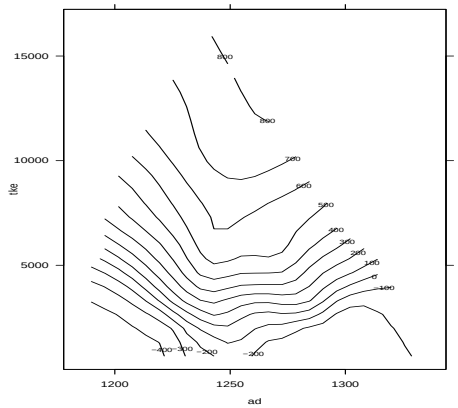
(a) Weight for DWD in D/H



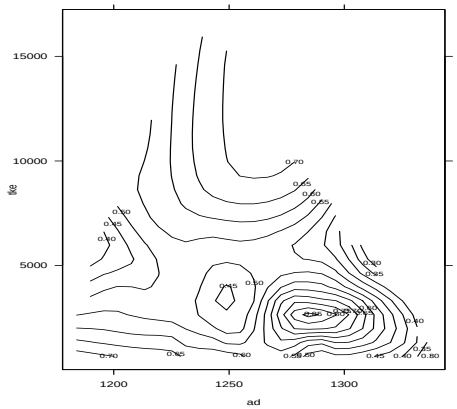
(b) Intercept in D/H



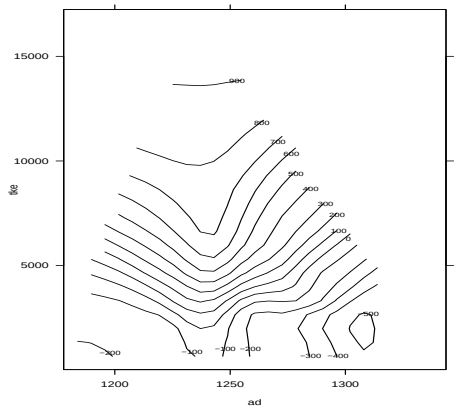
(c) Weight for DWD in D/M



(d) Intercept in D/M

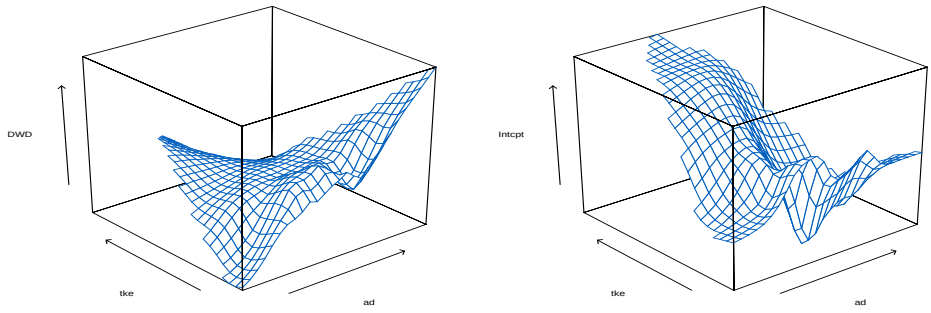


(e) Weight for HIRLAM in H/M



(f) Intercept in H/M

Figure 17: Contourplots for weights and the intercepts.



(a) Surface plot of $w_1(\mathbf{ad}, \mathbf{tke})$ (b) Surface plot of $w_0(\mathbf{ad}, \mathbf{tke})$

Figure 18: Surface plots for the DWD weight and the intercept.

By omitting the deep valley on the surface of the intercept, the relationship between the weight and the air density appears to be bell-shaped functions which fades out with increase in **tke**. Such a function can be difficult to formulate and implementing into the model would give complicated interpretation. The naive assumption is that **ad** does not affect the intercept but the intercept is a function of **tke**:

$$w_0(\mathbf{ad}, \mathbf{tke}) = v_0(\mathbf{tke}). \tag{36}$$

This assumption might be a bit crude but will make it easier to implement the estimated functions from (35) and (36) to a conditional parametric model:

$$\begin{aligned} \widehat{y}_c - \widehat{y}_2 &= v_0(\mathbf{tke}) + [v_{10}(\mathbf{tke}) + v_{11}(\mathbf{tke}) \cdot \mathbf{ad}]\,(\widehat{y}_1 - \widehat{y}_2) \\ &= v_0(\mathbf{tke}) + v_{10}(\mathbf{tke})(\widehat{y}_1 - \widehat{y}_2) + v_{11}(\mathbf{tke})(\widehat{y}_1 - \widehat{y}_2)\mathbf{ad} \\ &= v_0(\mathbf{tke}) + v_{10}(\mathbf{tke})z_1 + v_{11}(\mathbf{tke})z_2 \end{aligned} \tag{37}$$

where $z_1 = \widehat{y}_1 - \widehat{y}_2$ and $z_2 = (\widehat{y}_1 - \widehat{y}_2)\mathbf{ad}$. The model in (37) is now a modified CPM where the parameters are now only depending on one unknown variable instead of two, namely the turbulent kinetic energy.

Figure 19 shows how the weights in (37) change with **tke**. The same valley appears for the intercept, as in Figures 17 and 18, and the same test is performed as before to estimate sufficiency of the variance of the intercept. Table 6 shows the covariance matrix of the terms in (37) and reveals that the variance of the intercept is only fraction of the other variances and therefore the valley at **tke** around 3000 can be neglected. The opposite behavior of the parameters v_{10} and v_{11} is not surprising since they form the weight in the original model (31). From Figure 19 it can be assumed that the parameters, v_{10} and v_{11}, are linear functions of **tke** which changes slope when **tke** is approximately 9000.

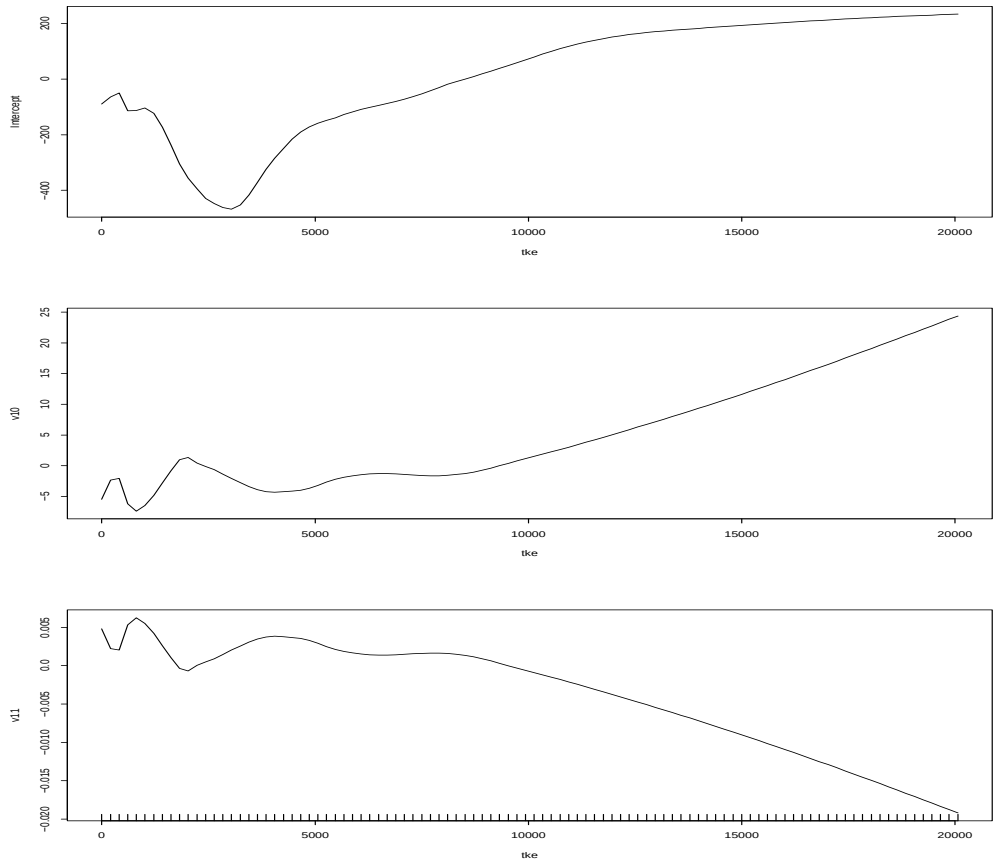


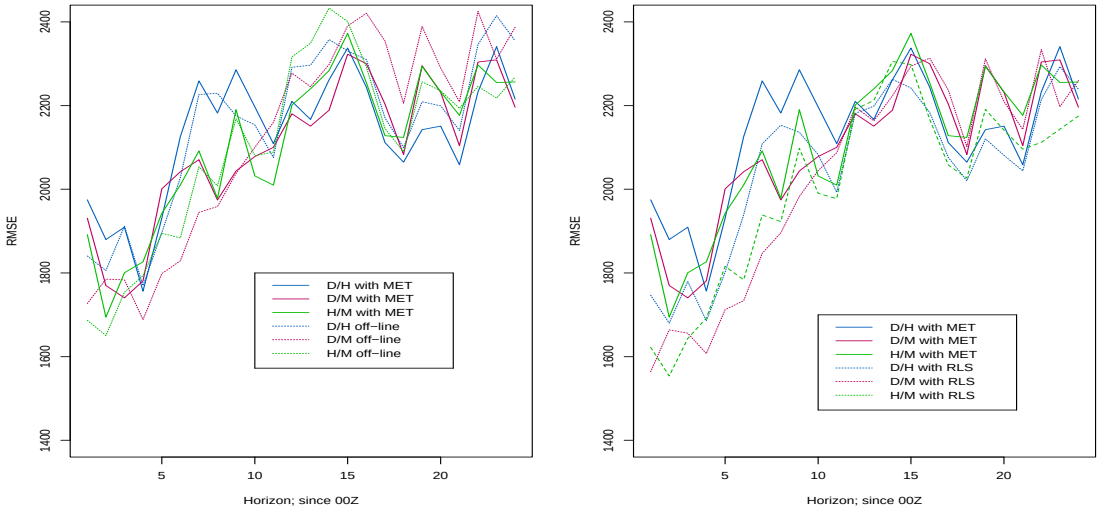
Figure 19: Coefficients in (37) as a functions of turbulent kinetic energy.

6.3 Comparison with foregoing methods

The performance for the conditional parametric model is compared with the RLS method and the offline model from Section 4.2. The performances for the surface estimations above are generated by insample RMSE and coefficient of determination, which was also performed for the offline estimation. In Figure 20(a) the forecasts generated by using MET variables are compared to the offline performance. For the first 12 prediction horizons the methods are performing quite alike except DWD/MM5. That specific forecasts is very different than the competing performance for the first 8 horizons, but thereof it

Table 6: Covariance matrix for the terms in (37).

Variance	Intercept	$v_{10}z_1$	$v_{11}z_2$
Intercept	$2.37259e4$	$9.76257e4$	$-7.741621e4$
$v_{10}z_1$	$9.76257e4$	$7.445344e7$	$-7.818895e7$
$v_{11}z_2$	$-7.741621e4$	$-7.818895e7$	$8.304898e7$



(a) MET vs. offline

(b) MET vs. RLS

Figure 20: Compare MET dependent forecasts with other performances

has lower RMSE. For larger horizons forecasts using MET variables are outperforming the offline model, estimated over the data set. On average is the improvement 1%, but considering not the shorter prediction horizons this improvement is closer to 2%.

With MET based forecasts outperforming the offline performance for large horizons it is interesting to compare it with the RLS performance. This is depicted in Figure 20(b). This comparison reveals that over the intermediate prediction horizons the MET dependent forecasts are very close to the RLS performance. For small horizons all the combined forecasts from RLS are significantly better. The difference is highest for the shorter horizons but then decreases until the intermediate horizons. For the largest horizons both DWD/HIRLAM and DWD/MM5 forecasts are performing similarly, but the difference between the MET dependent HIRLAM/MM5 forecast and corresponding RLS forecast increases. The performance of the conditional parametric model in (37) is depicted in Figure 21 in orange. It appears to be approaching the offline performance but with a slight improvement.

In Table 7 R^2 for MET dependent forecasts are compared to the coefficient of determination for the offline method and RLS method. From the table it can be concluded that the fit for MET based forecasts is not as good as for the other forecast methods. But the local regression, using MET forecasts as predictors for the weights, is an static procedure which used only a fraction of the data set to estimate a fitting point. The performance for the method can be improved by estimating the weights with adaptive estimation.

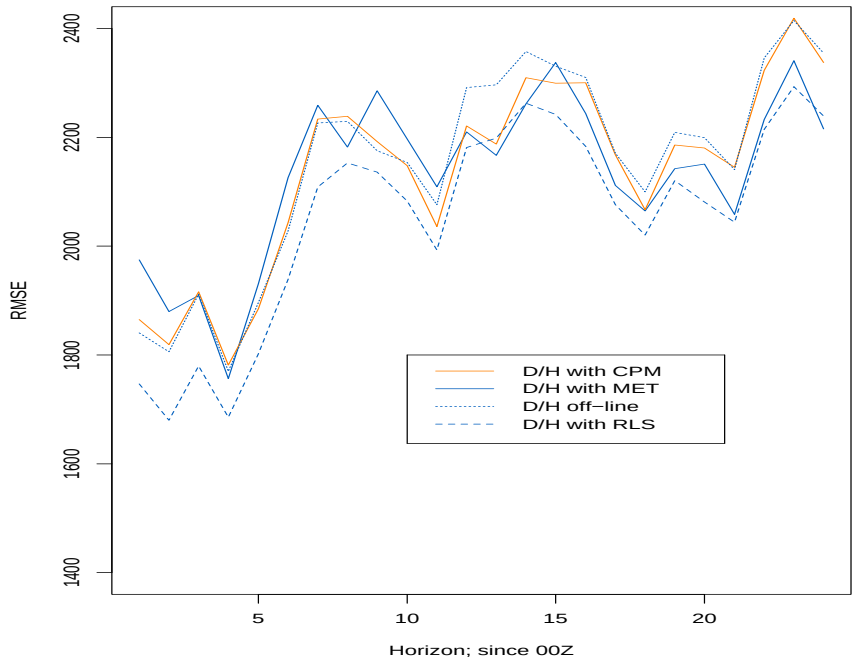


Figure 21: Performance of the modified model in (37) compared to other DWD/HIRLAM performances.

7 Conclusion and discussions

In the study the idea of combining wind power forecasts has been demonstrated to obtain more adequate performance for the power prediction. Various methods for combining

Table 7: Coefficient of determination (R^2) for combining forecasts. Comparing the MET dependent forecasts to the foregoing methods in this study. The results are shown for selected prediction horizons between 1 hour and 24 hours.

Combination		Prediction horizon						
		1	2	3	6	12	18	24
CPM	D/H	0.673	0.706	0.692	0.613	0.602	0.645	0.586
	D/M	0.714	0.761	0.772	0.687	0.634	0.672	0.635
	H/M	0.730	0.784	0.756	0.695	0.634	0.660	0.616
Offline	D/H	0.770	0.796	0.791	0.781	0.817	0.724	0.681
	D/M	0.807	0.827	0.829	0.821	0.847	0.758	0.702
	H/M	0.818	0.848	0.834	0.818	0.838	0.689	0.701
RLS	D/H	0.803	0.815	0.810	0.815	0.838	0.812	0.694
	D/M	0.861	0.840	0.854	0.869	0.855	0.817	0.726
	H/M	0.850	0.860	0.856	0.862	0.855	0.829	0.746

forecasts have been introduced where the linear regression model, including both a constant term and an weight restriction, is given a detailed description, for both offline and online procedures. If the constant is omitted in the linear model it becomes the equal to the minmum variance method, but comparison of these two methods reveals the importance of a constant in the model since there is a significant difference between these two methods in performance. The method of simple average for the weights in the combination is also applied and was the least performing method.

In advance to the linear model, the weights were fitted with local regression and improved the recursive least squares method when the bandwidth spanned about 50 days, which was the optimal sliding window (past observations) used in the recursive least squares method. However, selection of bandwidth has a tradeoff between variance and bias and the analysis concluded that weights estimated with bandwidth of 50 days in local regression as significantly greater variance than corresponding weight estimation with recursive least squares. The variance became the same when the bandwidth was increased to 80 days, which is closer the the 50 days of past values considered by the recursive least squares. However, by using data close to the fitting point, not only past observations, a phase error was detected in the recursive estimation. The local regression was considered to give an improved estimation of the weights, and was thus used to allocate the two meteorological forecasts, air density and turbulent kinetic energy, to generate the weights in the combined forecast, i.e. the linear model was extended to a conditional parametric model.

The weights depending on the meteorological forecasts gave similar results as the offline method for the shorter prediction horizons, but improved the offline method significantly for the larger prediction horizons. On average was the improvement 1%, but by only considering the larger prediction horizons (12-24) this improvement was very close to 2%. The improvement for larger prediction horizons was such that it became very close to the recursive least squares performance, that is when DWD was included in the combination.

Applying the air density and turbulent kinetic energy provided a smoothed surface for the weight estimations in combinations including the individual forecast DWD. The surface for the DWD weights in DWD/HIRLAM combinaton was further extended by considering the surface as a linear function of air density where the intercept and the slope were dependent on turbulent kinetic energy. This was implemented in the linear model and resulted in the preformance improvement compared to the offline procedure. Though the difference is not statistically significant is it visual and could be further improved by adopting a recursive approach for the conditional parametric model.

Acknowledgements

The project is sponsored by the Danish utilities PSO fund (PSO2004 / FU5766) which is hereby greatly acknowledged. Also Danmarks Meteorologiske Institut (*DMI*) is acknowledged for providing the meteorological data set used in this study.

References

Clemen, R. T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5 (4), 559–583.

Cleveland, W. S., Dec 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74 (368), 829–836.

Cleveland, W. S., 1994. Multivariate Analysis and Its Applications. Vol. 24 of IMS Lecture Note-Monograph Series. Hayward, California, Ch. Coplots, nonparametric regression, and conditionally parametric fits, pp. 21–36.

Cleveland, W. S., Devlin, S. J., Sep 1988. Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83 (403), 596–610.

de Menezes, L. M., Bunn, D. W., Taylor, J. W., 2000. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 120 (1), 190–204.

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. Journal of the Royal Statistical Society, B 55 (4), 757–796.

Holden, K., Peel, D., 1989. Unbiasedness, efficiency and the combination of economic forecasts. Journal of Forecasting 8, 175–188.

Madsen, H., 2001. Time Series Analysis, 1st Edition. IMM,DTU.

Madsen, H., Holst, J., 2000. Modelling Non-linear and Non-stationary Time Series. IMM,DTU.

Madsen, H., Nielsen, H. A., Nielsen, T. S., 2005a. A tool for predicting the wind power production of off-shore wind plants. In: Proceedings of the Copenhagen Offshore Wind Conference & Exhibition.

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., Nielsen, T. S., 2005b. Standardizing the performance evaluation of short-term wind power prediction models. Wind Engineering 29 (6), 475–489.

Montgomery, D. C., Runger, G. C., 2002. Applied statistics and probability for engineers, 3rd Edition. John Wiley & Sons, USA.

Nielsen, H. A., Nielsen, T. S., Joensen, A. K., Madsen, H., Holst, J., 2000. Tracking time-varying-coefficient functions. International Journal of Adaptive Control and Signal Processing 14, 813–828.

Sánchez, I., June 2006. Adaptive combination of forecast with application to wind energy forecast.

Thordarson, F. Ø., 2007. Optimal combined wind power forecasts using exogenous variables. Master's thesis, IMM-DTU, Kgs. Lyngby.