Optimal combined wind power forecasts using exogeneous variables

Fannar Örn Thordarson

Kongens Lyngby 2007

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-M.Sc: ISSN 1601-233X

Summary

The aim of combining forecasts is to reduce variation from observed values by compositing two or more forecasts, which predict for the same event at the same time. Many methods have developed since the problem was presented, varying from a method of equal weights to more complex methods e.g. state space. Despite the complexity a linear model of the combination appears to be the most favorable 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) estimate weights by local regression and compare with RLS and minimum variance methods, which are well known procedures when 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 when it outperforms the RLS estimation which is also considered. Furthermore, the meteorological forecasts from DMI-HIRLAM are inspected and the air density (**ad**) and the turbulent kinetic energy at pressure level 38 (**tke**) are used as regressors for locally fitting the weights into the linear model.

The results in this report show that using the meteorological information to estimate the weights does not outperform the RLS method but does give reasonable fit, which can be elevated by further analysis.

Preface

This thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the degree, Master of Science in Engineering.

The project was carried out in the period from February 1st 2006 to January 2nd 2007.

The subject of the thesis is combined wind power forecasts using informations from meteorological forecasts.

Lyngby, January 2007

Fannar Örn Thordarson

Acknowledgements

I thank my supervisors Henrik Madsen and Henrik Aalborg Nielsen for their guidance throughout this project.

Contents

Chapter 1

Introduction

1.1 Background

Where more then one forecast for some event at the same time is available, it can be attractive procedure to combine the forecasts. By combining the independent informations included in every individual forecast are gathered and more accurate forecast can be accomplished. The application of combining wind power forecasts for certain wind power plant is appealing procedure where 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 so highly unstable that a good forecast is needed. With several such forecasts, more accurate forecast can be acquired by combining.

Many combining methods have been developed since it was first introduced as objective procedure, ranging from simple average of the constituent forecasts to far more complex mode like state space. Despite all this methodology for compositing, adoption of the linear regression has always been the most attractive procedure.

1.2 Aim of thesis

In this presentation the weights are tracked over time for 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 of this presentation 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 1.1 shows the flow of combining forecasts and also how the informations 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 any past data.

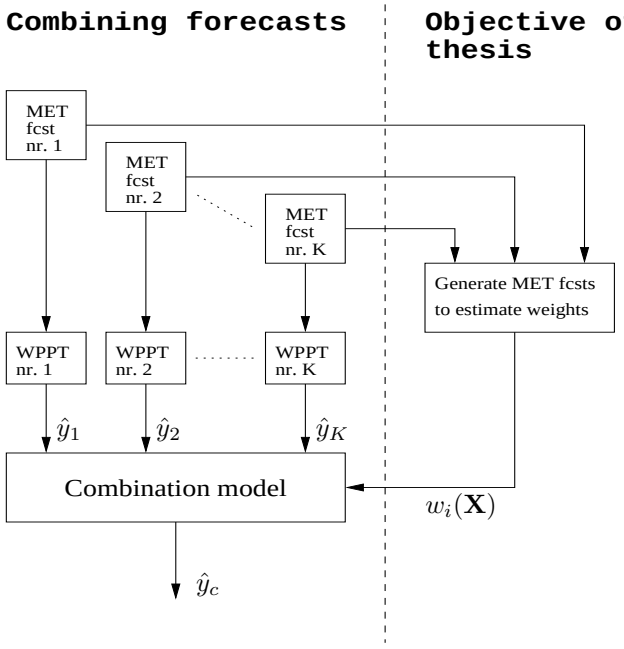


Figure 1.1: Block diagram of the process of combining wind power forecasts. To the left of the dashed line the flow for combining WPPT forecasts is described, but the right side shows the black box model for the weights with the MET forecasts as input.

1.3 Outline of thesis

The thesis is divided in three main elements where first, chapter 2 and 3 take care of the methodology used in the analysis; second, the data is introduced in detail in chapter 4; and third, the analysis in chapters 5 to 7. Finally the thesis are concluded in chapter 8. More detailed description of the chapters is listed below:

Chapter 2 describes the methods used to combine forecasts with the linear model. Also short discussion about uncertainty when forecasts are aggregated and some performance measures are presented.

Chapter 3 introduces varying coefficient-functions which are generated when the weights are defined as functions of some meteorological variables.

Chapter 4 gives a quite detailed description about the data used in the analysis.

Chapter 5 includes the WPPT forecasts combined with RLS and minimum variance methods. In the RLS estimation there is an issue of involving intercept in the linear model which gets some attention in one section along with a restriction on the forecast weights.

Chapter 6 describes local regression as it is used to fit the weights in the combination. These weights are compared with the estimates from RLS w.r.t. the bandwidth.

Chapter 7 illustrates how the meteorological forecasts can be used to estimate weights in combined forecast. In the analysis the varying coefficient-functions are applied to explain the dependencies.

Chapter 8 concludes on the thesis and includes a section about further works.

The analysis in the thesis was mainly carried out in the statistical software S-PLUS, but also MATLAB was considered.

Chapter 2

Methods of combining forecasts

2.1 Introduction

The reason for combining forecasts for some event, where two or more individual forecasts are available for the same event, is to reach the goal of improved forecast. Before the seminal work of Bates and Granger [10] some attempts were on revealing which forecast is the best, then accept it and discard the others. This procedure might have some justification but if the objective is to observe as good forecast as possible, independent informations each individual forecast contains might be important.

From the late sixties when Bates and Granger publiced there work, the methodology of combining forecasts has developed within many applications. In [2] Clemen publiced a bibliographical review of forecast combining, but since then the literature has grown substantially. Clemen divides the literature regarding statistics, management, psychology and many more. The main interest in this presentation concerns the statistical methods used to reach the objective of minimizing the variance error.

Before discussing the methods of combining forecasts the combination model is introduced in 2.2. Sections 2.3 to 2.7 introduce various methods of combining

forecasts and in 2.8 adaptive estimation is described. Little discussion about uncertainty in combined forecasts is in 2.9 and the last section of the chapter,2.10, list several accuracy criterions often used as performance measurements for the combined forecasts.

2.2 Combination model

The variable of interest is the one we want to predict and at time t it is denoted as y_t. When applying to wind power prediction, the variable of interest is the actual wind energy production. Let $\hat{y}_{i,t|t-h}$ be the i-th individual forecast at time t with available information at time $t-h$. The prediction error between the production and the i-th individual forecast is

$$e_{i,t|t-h} = y_t - \hat{y}_{i,t|t-h} \tag{2.1}$$

where $i = 1, \ldots, K$. The lead h also catagorizes the error terms and has to be taken into account. A linear combination of the competing forecasts is then formulated as

$$\hat{y}_{c,t|t-h} = w_{0,t}(h) + \sum_{i=1}^{K} w_{i,t}(h)\hat{y}_{i,t|t-h} \tag{2.2}$$

where $w_{i,t}(h)$ is the weight given to model i at time t. The weights are also depending on its horizon h but for convenience it is omitted throughout this presentation. The $w_{0,t}$ term represents the constant term in the linear model. The combination model can be written as

$$y_t = \hat{y}_{c,t} + e_{c,t} = w_{0,t} + \sum_{i=1}^{K} w_{i,t}\hat{y}_{i,t} + e_{c,t}. \tag{2.3}$$

which gives the prediction error for the combined foreacast

$$e_{c,t} = y_t - \hat{y}_{c,t}. \tag{2.4}$$

Due to accuracy the squared errors are minimized w.r.t. the weights to gain optimal coefficients for the constituent forecasts. The method of least squared errors is illustrated in section 2.6.1

The linear formulation can be written as

$$\hat{y}_{c,t} = \mathbf{w_t^T}\hat{\mathbf{y}}_\mathbf{t} \tag{2.5}$$

where

$$\hat{\mathbf{y}}_\mathbf{t} = (\hat{y}_{1,t}, \hat{y}_{2,t}, \ldots, \hat{y}_{K,t})^T \tag{2.6}$$

is a vector of K individual forecasts to be combined at time t and

$$\mathbf{w_t} = (w_{1,t}, w_{2,t}, \ldots, w_{K,t})^T \tag{2.7}$$

is a vector of linear weights assigned to each individual forecast at time t. It should be noted that when constant is included, it is counted in the vector of weights and with unity in the vector of individual forecasts. This presentation is simply a vector notation of (2.2).

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

$$\sum_{i=1}^{K} w_{i,t} = 1. \tag{2.8}$$

The issue of adopt this restriction when combining forecasts has been discussed from the beginning. In section 5.3 this matter is discussed along with including constant in the combining.

2.3 Simple average method

This methods is the most 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

$$w_i = \frac{1}{K} \tag{2.9}$$

where the index i is reference to inidividual forecast i in the combined forecast.

In [2] Clemen states the question of why the simple average work so well, but the method has been a conclusive method to use in quite many studies. In [**?**] Gunter identified analytically the conditions for majority of the simple average over the optimal and regression methods explained below. But in [5] Granger had concluded that using equal weights in combining are useful when information components, to be combined, are common and independent.

2.4 Outperformance method

Each individual weight is interpreted as the probability that its respective forecast will perform the best in the next occation. Each probability is estimated as the fraction of occurrences in which its respective forecasting model has performed the best in the past. The forecast combination is developed as $\hat{y}_c = \mathbf{p}^T\hat{\mathbf{y}}$ where p is a vector of probabilities for individual forecasts.

This method was proposed in [1] where it is shown how subjective probabilities can be assigned over a set of forecasting models and updated when the forecast realizations become known. In [15] a Bayesian framework is developed to encode subjective knowledge about the information sources in order to combine point forecasts.

The method of Bayesian approach is not applied in this study.

2.5 Optimal method

The combination method is denoted as *optimal* when the individual weights are calculated to minimize the squared residuals of the combination where the assumption about unbiasedness for each individual forecast is made. The vector of combining weights, $\mathbf{w}$, is determined by the formula

$$\mathbf{w^s} = \frac{\mathbf{S^{-1}u}}{\mathbf{u^T S^{-1} u}} \tag{2.10}$$

where $\mathbf{u}$ is the $n \times 1$ unit vector and $\mathbf{S}$ is the $n \times n$ covariance matrix of the forecast errors. The problem with this approach is that the covariance matrix $\mathbf{S}$ has to be properly estimated.

Applying the optimal method, more efficiency could be gained if the forecast errors of the individual predictions were treated as independent. This implies that the elements of $\mathbf{S}$ in equation (2.10) are restricted to be the diagonal terms of the covariance matrix. When assuming independence, the weights in the combination are obtained by

$$\mathbf{w^v} = \frac{\mathbf{V^{-1}u}}{\mathbf{u^T V^{-1} u}} \tag{2.11}$$

where $\mathbf{V} = \text{diag}(\mathbf{S})$. The elements of the diagonal aggregate to unity, so the optimal method offers restriction on the weights.

In practice the covariance matrix is often time-varying which implies that the weights in (2.10) are estimated adaptively. Appropriate initial weights for such an approach is to use equal weights as illustrated in equation (2.9) where the inverse of the weights are the diagonal terms of the initial covariance matrix $\mathbf{S}$. The adaptive approach for combining will be explained in section 2.8.

2.6 Regression method

The classical regression model is used to describe a static relation between a dependent variable, which in case of combining are the observations, and one or more predictor variables, the individual forecasts. In [6] Granger and Ramanathan combined forecasts as an unrestricted least squares regression with an intercept and showed that their method outperformed the optimal method if the individual forecasts were biased. Their conclusion encouraged many researchers to focus on the area of combining with regression, where the issues of constant and the sum-to-unity restriction were applied in the combined forecasts. This is of concern in section 5.3.

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

$$y_t = g(\hat{\mathbf{y}}_t, t; \mathbf{w}) + \varepsilon_t \tag{2.12}$$

where $g(\hat{\mathbf{y}}_t, t; \mathbf{w})$ is a known mathematical function of the independent individual forecasts $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_N)^T$, but the weights $\mathbf{w} = (w_1, \ldots, w_N)^T$ are unknown. ε_t is a random variable with $E[\varepsilon_t] = 0$ and $V[\varepsilon_t] = \sigma_\varepsilon^2$. It is also assumed that $cov[\varepsilon_{ti}, \varepsilon_{tj}] = \sigma\mathbf{\Sigma}_{ij}$ and ε_t and $\hat{\mathbf{y}}_t$ are independent. For the regression model in (2.12) for y_t^c given $\hat{\mathbf{y}}_t$ the following holds:

$$E[y_{c,t}|\hat{\mathbf{y}}_t] = g(\hat{\mathbf{y}}_t, t; \mathbf{w}) \tag{2.13}$$

$$V[y_{c,t}|\hat{\mathbf{y}}_t] = \sigma^2 \tag{2.14}$$

$$cov[y_{c,ti}, y_{c,tj}|\hat{\mathbf{y}}_t] = \sigma\mathbf{\Sigma}_{ij} \tag{2.15}$$

In this presentation the attention is on a model where the independent variables are the individual forecasts used in a combined forecast and are therefore fixed.

The general linear model is a special case of the regression model where the estimated response is a linear function on its parameters:

$$y_{c,t} = \hat{\mathbf{y}}_t^T \mathbf{w}_t + \varepsilon_t. \tag{2.16}$$

The properties of this structure is the same as mentioned above for the regression model where the parameters of the model are unknown. These parameters need

to be properly estimated. Several methods are available, but two methods are more exploited for estimations in the general linear model, the least squares estimation and the maximum likelihood estimation.

2.6.1 Least Squares (LS) estimates

To obtain weights for the most adequate model some loss function L is minimized. For the combined forecast to be competetive with the individual forecasts its loss function has to have lower magnitude than the individual loss functions. The solution to the combination problem is a vector of weights, $\mathbf{w}_t = (w_1, ..w_K)^T$, such that it minimizes the loss function $L(e_c)$ in a way that

$$L\left(e_c\right) \leq \min_i \left\{L\left(e_i\right)\right\}. \tag{2.17}$$

The loss function for the combination can be assumed to be the least squares function, $L(e_c) = E[(e_c)^2]$. Then the LS method is used to estimate the weights in the combination.

The LS method estimates the weights such that the total squared error is minimized,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} S(\mathbf{w}) \tag{2.18}$$

where the function to minimized is the quadratic loss function noted as

$$S(\mathbf{w}) = \sum_{i=1}^{N} e_i^2 = \mathbf{e^T e} = (y - \hat{\mathbf{y}}^T\mathbf{w})^T(y - \hat{\mathbf{y}}^T\mathbf{w}). \tag{2.19}$$

The minimization problem in (2.19) is solved with differentiation on the loss function w.r.t. the weighting vector. By equalling the derivative to zero the objective estimator for the quadratic loss is observed. The vector $\hat{\mathbf{w}}$ is an estimator of the weights in $\mathbf{w}$ that minimizes the sum of squared residuals,

$$\hat{\mathbf{w}} = (\hat{\mathbf{y}}^{\mathbf{T}}\hat{\mathbf{y}})^{-\mathbf{1}}\hat{\mathbf{y}}^{\mathbf{T}}y. \tag{2.20}$$

The equation in (5.7) 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

$$\hat{\mathbf{w}} = (\hat{\mathbf{y}}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\hat{\mathbf{y}})^{-\mathbf{1}}\hat{\mathbf{y}}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}y \tag{2.21}$$

where $\hat{\mathbf{y}}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\hat{\mathbf{y}}$ has full rank.

The static LS method is very useful when working with not too large data sets, but when more informations are added to the data when observations become available, the model parameters need to be updated recursively. On-line methods are presented in section 2.8.

2.6.2 Maximum Likelihood (ML) estimates

It is assumed that the statistical model for the observations $Y_1, \ldots, Y_n$ is given by a family of joint densities, $\{f_Y(y_1, \ldots, y_n; w)\}_{w \in W}$. For convenience, the observation set is defined as $\mathbf{y} = (y_1, \ldots, y_n)$.

Definition 2.1 (Likelihood Function) For the given observations $\mathbf{y}$, the likelihood function for the estimates w is the function

$$L(w; \mathbf{y}) = c(y_1, \ldots, y_n) f_Y(y_1, \ldots, y_n; w) \qquad w \in W \tag{2.22}$$

where $c(y_1, ..y_n)$ is a constant, given the observations. The likelihood function is therefore the joint probability density for the actual observations considered as a function of w.

In the definition of the likelihood function the parameters w are unknown, but these parameters are wanted to be such that the likelihood function is maximized. The maximum likelihood estimator (MLE), denoted be $\hat{w}$, is the value of w that maximizes $L(w; \mathbf{y})$.

Definition 2.2 (ML-estimates) Given the observation $\mathbf{y}$ the ML estimate is a function $\hat{w}(\mathbf{y})$ such that

$$L(\hat{w}; \mathbf{y}) = \sup_{w \in W} L(w; \mathbf{y}) \tag{2.23}$$

The function $\hat{w}(\mathbf{Y})$ over the sample space of observations Y is called a ML estimator.

The maximum of $l(w; \mathbf{y}) = \log(L(w; \mathbf{y}))$ occurs at the same place as the maximum of $L(w; \mathbf{y})$ and in practice it is more convenient to work with the log-likelihood function. With respect to the definition of MLE, when the supremum is attained at an interior point, the estimates can be obtained by solving

$$\frac{\partial}{\partial w} l(w; \mathbf{y}). \tag{2.24}$$

The maximum likelihood estimator is considered in the general linear model, $\mathbf{y} = \hat{\mathbf{y}}^{\mathbf{T}}\mathbf{w} + \boldsymbol{\epsilon}$. The ML estimates are now based on the assumption that the

residuals $\boldsymbol{\epsilon}$ are normally distributed, and thus $\mathbf{y}$ as well. If $\mathbf{y}$ is considered to be a random vector of n observations where $\mathbf{y} \in N(\hat{\mathbf{y}}^{\mathbf{T}}\mathbf{w}, \sigma^2\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ assumed to be known, the maximum likelihood estimator for $\mathbf{w}$ is

$$\hat{\mathbf{w}} = (\hat{\mathbf{y}}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\hat{\mathbf{y}})^{-1}\hat{\mathbf{y}}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{Y}. \tag{2.25}$$

This is the same as for the weighted least squares estimator. Thus it can be concluded that under the assumption of normality, the maximum likelihood estimator is equivalent to the least square estimator.

2.7 State Space

State space models are very useful for describing non-stationary or time-varying systems. For stochastic systems, the state vector at a given time contains all information available for the future evaluation of the system.

The major advantage of combining forecasts with the state space approach over the optimal and regression methods, is that no estimation of forgetting factor is needed. The state space model for combining can be interpreted as

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \mathbf{e}_{1,t}, \tag{2.26}$$
$$y_t = \mathbf{w}_t\hat{\mathbf{y}}_t + \mathbf{e}_{2,t}, \tag{2.27}$$

where (2.26) is *system equation* and (2.27) the *observation equation*. $\mathbf{w}_t$ is a n-dimensional stochastic state vector, y_t is a observation at time t and $\hat{\mathbf{y}}_t$ is a vector of individual forecasts at time t. In general, $\{\mathbf{e}_{1,t}\}$ and $\{\mathbf{e}_{2,t}\}$ are stochastic vectors where

$$E[\mathbf{e}_{1,t}] = E[\mathbf{e}_{2,t}] = 0 \tag{2.28}$$
$$C[\mathbf{e}_{\cdot,t}, \mathbf{e}_{\cdot,s}] = \begin{cases} V[\mathbf{e}_{\cdot,t}] = \boldsymbol{\Sigma}_{\cdot,t} & \text{for} \quad s = t, \\ V[\mathbf{e}_{\cdot,t}] = 0 & \text{for} \quad s \neq t \end{cases}, \tag{2.29}$$
$$C[\mathbf{e}_{1,t}, \mathbf{e}_{2,t}] = 0 \qquad \forall \quad s, t. \tag{2.30}$$

The Kalman filter yields the optimal reconstruction and prediction of the state space $\mathbf{w}_t$ given the observations of y_t in the state space model. The optimal linear reconstruction $\hat{\mathbf{w}}_{t|t}$ and corresponding variance of the state space system is

$$\hat{\mathbf{w}}_{t|t} = \hat{\mathbf{w}}_{t|t-1} + \mathbf{K}_t\left(y_t - \hat{\mathbf{y}}\hat{\mathbf{w}}_{t|t-1}\right), \tag{2.31}$$
$$\boldsymbol{\Sigma}^{ww}_{t|t} = \boldsymbol{\Sigma}^{ww}_{t|t-1} - \mathbf{K}_t\boldsymbol{\Sigma}^{yy}_{t|t-1}\mathbf{K}_t^T = \boldsymbol{\Sigma}^{ww}_{t|t-1} - \mathbf{K}_t\hat{\mathbf{y}}_t\boldsymbol{\Sigma}^{ww}_{t|t-1}, \tag{2.32}$$

with the Kalman gain

$$\mathbf{K}_t = \mathbf{\Sigma}^{ww}_{t|t-1}\hat{\mathbf{y}}^T_t \left(\mathbf{\Sigma}^{yy}_{t|t-1}\right)^{-1}. \tag{2.33}$$

The formula denotes how information from past observations are combined with a new observation y_t to improve the estimates of the weights, $\mathbf{w}$. It is quite clear from above that in order to calculate the reconstruction and corresponding variance, calculation of one-step prediction of $\mathbf{w}_t$ and its corresponding variance is needed. The prediction $\hat{\mathbf{w}}_{t+1|t}$ of the system is obtained with

$$\hat{\mathbf{w}}_{t+1|t} = \hat{\mathbf{w}}_{t|t}, \tag{2.34}$$
$$\mathbf{\Sigma}^{ww}_{t+1|t} = \mathbf{\Sigma}^{ww}_{t|t} + \mathbf{\Sigma}_1, \tag{2.35}$$
$$\mathbf{\Sigma}^{yy}_{t+1|t} = \hat{\mathbf{y}}_t \mathbf{\Sigma}^{ww}_{t+1|t}\mathbf{f}^T_t + \mathbf{\Sigma}_2, \tag{2.36}$$

where the initial conditions are

$$\hat{\mathbf{w}}_{1|0} = E\left[\mathbf{w}_1\right] = \boldsymbol{\mu}_0 \tag{2.37}$$
$$\mathbf{\Sigma}^{ww}_{1|0} = V\left[\mathbf{w}_1\right] = \mathbf{V}_0. \tag{2.38}$$

This method is not applied in the following studies.

2.8 Adaptive combination of forecasts

In many applications it is necessary to implement a procedure who work online. That means when new informations become available they are added to the data set and the estimation is updated. This recursion allows the parameters to change with the time which implies time-varying estimation.

Two methods of adaptation are described, the exponentially weighted moving average and recursive least squares method. The former method is applied to adaptively estimate the minimum variance method (optimal method), while the latter updates the regression model when new observations become available.

2.8.1 Exponentially Weighted Moving Average (EWMA)

The new observation arriving the data set every time it becomes available, will evantually loose its importance if the entire data set is used every time the parameters are estimated. Therefore the process need to have a sliding window where the estimation only uses part of past observations. It would be reasonable to give more weight to the most recent observation and less weight to the

observations in the past.

Within the moving average, some λ-value is estimated which gives the importance of the new data against the past data. This value is interpret as the fraction of the present average, given the new observation, to estimate the average in next time step. A moving average of a process can be interpreted as

$$\mu_t = \lambda\mu_{t-1} + (1-\lambda)y_t \tag{2.39}$$

where μ_t is the moving average at time t and y_t is the observed value at time t. When the value of λ is constant, it is called *the forgetting factor* and is $\lambda \in [0,1]$. Using the moving average in (2.39), an initial estimate for μ_0 is needed. The average of some recent data can be used, but the influence of the initial guess will decay rapidly since

$$\begin{aligned}\mu_t &= \lambda^2\mu_{t-2} + (1-\lambda)y_{t-1}) + (1-\lambda)y_t \\ &= \lambda(\lambda\mu_{t-2} + (1-\lambda)\left[\lambda y_{t-1} + y_t\right] \\ &\vdots \\ &= \lambda^N\mu_0 + (1-\lambda)\left[\sum_{i=1}^{N}\lambda^{N-i}y_i\right].\end{aligned} \tag{2.40}$$

Alternatively the first observation can be used as the initial value.

It can be seen from (2.40) that when informations get older, the corresponding forgetting factor decreases drastically. The moving average in (2.39) is thus called exponential smoothing since the weights on past data decay exponentially.

The use of foregetting factor can be applied to the optimal method to generate an adaptive estimation for the weights. With appropriately defined λ, the adaptive estimator of $\mathbf{S}_t$ can be given with

$$\hat{\mathbf{S}}_t = (1-\lambda)\mathbf{e}_t\mathbf{e}_t^T + \lambda\hat{\mathbf{S}}_{t-1}. \tag{2.41}$$

The values of a vector $\mathbf{V}_t$ in (2.11) are estimated with $\hat{\mathbf{V}}_t = \mathrm{diag}(\hat{\mathbf{S}}_t)$. The initial matrix $\hat{\mathbf{S}}_0$ can not be considered to be zero, it is evaluated from some past data or some part of the data set. An appropriate value for λ is between 0.95 and 0.99.

2.8.2 Recursive Least Squares (RLS)

In many applications it is is necessary to implement a regression model as (2.12) where the model parameters are updated recursively as the available observations are added to the data set. Wind power systems usually work on-line in

an environment with minimum user interference for long periods of time. With that in mind some time evolution in the dynamic properties of the predictor should be allowed, which indicates that the weights of the combination should be time-varying.

The recursive approach for the parameter estimations is derived at time t with given information to time $t-1$ where $\mathbf{R}_t$ and $\mathbf{g}_t$ can be written recursively as

$$\begin{aligned} \mathbf{R}_t &= \mathbf{R}_{t-1} + \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t^T && (2.42) \\ \mathbf{g}_t &= \mathbf{g}_{t-1} + \hat{\mathbf{y}}_t y_t. && (2.43) \end{aligned}$$

where the $\hat{\mathbf{y}}_t$ is the vector of individual forecasts defined in (2.6) at time t given data at $t-1$. With these updates the new parameter estimates, $\hat{\mathbf{w}}_t$, can be found by inserting the updated formulas into (5.7). Since the LS method is now applied for the combination where the coefficients are the weighted estimator $\hat{\mathbf{w}}_t$, we assume that $\mathbf{w}_t = \theta_t$ and thus

$$\begin{aligned} \hat{\mathbf{w}}_t &= \mathbf{R}_t^{-1} \mathbf{g}_t && (2.44) \\ &= \mathbf{R}_t^{-1} \left[\mathbf{g}_{t-1} + \hat{\mathbf{y}}_t Y_t \right] && (2.45) \\ &= \mathbf{R}_t^{-1} \left[\mathbf{R}_{t-1} \hat{\mathbf{w}}_{t-1} + \hat{\mathbf{y}}_t y_t \right] && (2.46) \\ &= \mathbf{R}_t^{-1} \left[\mathbf{R}_t \hat{\mathbf{w}}_{t-1} - \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t^T \hat{\mathbf{w}}_{t-1} + \hat{\mathbf{y}}_t y_t \right] && (2.47) \\ &= \hat{\mathbf{w}}_{t-1} + \mathbf{R}_t^{-1} \hat{\mathbf{y}}_t \left[y_t - \hat{\mathbf{y}}_t^T \hat{\mathbf{w}}_{t-1} \right]. && (2.48) \end{aligned}$$

The updates in equations (2.42) and (2.48) for $\mathbf{R}_t$ and $\hat{\mathbf{w}}_t$ respectively, are referred as the *RLS method* for dynamical models. In order to avoid the inversion of $\mathbf{R}_t$ in each step, another notation is used

$$\mathbf{P}_t = \mathbf{R}_t^{-1} \tag{2.49}$$

By using the *matrix inversion rule*

$$\left[\mathbf{A} + \mathbf{BCD} \right]^{-1} = \mathbf{A}^{-1} \mathbf{B} \left[\mathbf{DA}^{-1} \mathbf{B} + \mathbf{C}^{-1} \right]^{-1} \mathbf{DA}^{-1} \tag{2.50}$$

where $\mathbf{R}_{t-1} = \mathbf{A}$, $\hat{\mathbf{y}}_t = \mathbf{B} = \mathbf{D}$ and $\mathbf{C} = \mathbf{I}$, the updating for $\mathbf{P}_t$ becomes

$$\mathbf{P}_t = \mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1} \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t^T \mathbf{P}_{t-1}}{1 + \hat{\mathbf{y}}_t^T \mathbf{P}_{t-1} \hat{\mathbf{y}}_t}. \tag{2.51}$$

In the RLS procedure above can only be considered when the weights are being constants in time. From previous discussion the coefficients are wanted to be time-varying. This is feasible if the RLS estimation minimizes the *weighted least squares estimator*, $\hat{\mathbf{w}}_t = \arg\min S_t(\mathbf{w})$, where $S_t(\mathbf{w})$ denotes the quadratic loss function

$$S_t(\mathbf{w}) = \sum_{s=1}^{t} \beta(t,s) \left(y_s - \hat{\mathbf{y}}_s^T \mathbf{w} \right)^2 \tag{2.52}$$

where $\{\beta(t,s)\}$ is assumed to be a sequence of weighting constants, expressed as

$$\beta(t,s) = \lambda(t)\beta(t-1,s) \qquad 1 \leq s \leq t-1 \tag{2.53}$$

$$\beta(t,t) = 1. \tag{2.54}$$

The quantity $\beta(t,s)$ is the weight given to the s-th residual in the quadratic loss function at time t. From (2.53) it is quite clear that the deviation between every two analog β's gives some value on λ relative to the time index. This implies that

$$\beta(t,s) = \prod_{j=s+1}^{t} \lambda(j) \tag{2.55}$$

where $0 < \lambda(j) \leq 1$. This procedure reduces the importance of old data, the smaller value of $\lambda(j)$ leads to lower influence of past data.

The solution to the weighted least squares problem is found with (2.44) where

$$\mathbf{R}_t = \sum_{s=1}^{t} \beta(t,s)\hat{\mathbf{y}}_s\hat{\mathbf{y}}_s^T \tag{2.56}$$

$$\mathbf{g}_t = \sum_{s=1}^{t} \beta(t,s)\hat{\mathbf{y}}_s y_s. \tag{2.57}$$

The RLS procedure with forgetting can now be found to be

$$\hat{\mathbf{w}}_t = \hat{\mathbf{w}}_{t-1} + \mathbf{P}_t\hat{\mathbf{y}}_t\left[y_t - \hat{\mathbf{y}}_t^T\hat{\mathbf{w}}_{t-1}\right] \tag{2.58}$$

$$\mathbf{P}_t = \frac{1}{\lambda(t)}\left[\mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t^T\mathbf{P}_{t-1}}{\lambda(t) + \hat{\mathbf{y}}_t^T\mathbf{P}_{t-1}\hat{\mathbf{y}}_t}\right] \tag{2.59}$$

In order to obtain the recursive estimation, it is important to provide an appropriated sequence of $\lambda(j)$'s for the performance of this adaptive process.

If $\lambda(j) = \lambda$ (constant), then λ is called forgetting factor. The loss function in (2.52) is then weighted exponentially as $\beta(t,s) = \lambda^{t-s}$ which gives the effective number of observations as

$$N_t = \sum_{i=0}^{\infty} \lambda^i = \frac{1}{1-\lambda} \tag{2.60}$$

Most applications use constant forgetting factor which ranges within 0.95 and 0.999.

2.8.3 Predicting for large horizons

To be able to use the *RLS-algorithm* for a larger predictions than 1-step ahead, the pseudo prediction error is used and is defined as

$$\tilde{y}_{t|t-h}^{pseudo} = y_t - \hat{y}_{t-h}^T \hat{w}_{t-1} \tag{2.61}$$

The k-step ahead prediction is calculated by

$$\hat{y}_{t+h|t} = \hat{y}_{t+h}^T \hat{\theta}_t \tag{2.62}$$

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 within each horizon, one step back means the last predicted value for a horizon. Therefore is the notation h, indication of the prediction horizon, omitted throughout the thesis.

2.9 Uncertainty in combined forecasts

Through the history of combining forecasts, the focus has mainly been on accuracy of the combination by assessing some criterion of different forecasting methods. Today a good point estimate is no longer sufficient where uncertainty and risk analysis are required, f.ex in business planning models. Of all the literature available about combined forecasts it is surprising how small fraction of it deals with this issue.

In [4] Menezes and Bunn try to formulate the uncertainty of combined forecast by specifying the forecast error distribution. They conclude that with respect to both shape of the forecast error distribution and corresponding stochastic behaviour, the forecast error variances should not be the only performance measure on combining. In [16] Taylor and Bunn estimated the predictive distribution using quantile regression. They concluded that in theory the quantile regression is inefficient due to correlation, but they encouraged researchers to investigate uncertainty in combined forecasts by using quantile regression.

2.10 Measurement of performance

There are various methods used as a measure of performance to choose an adequate model. Most of the measurements aim at minimizing the residuals

between a fit and corresponding observations. An often used model residual measure is the mean square error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N} e_i^2. \tag{2.63}$$

through the history of combining forecasts, this measure has been very popular. By taking the square root of MSE, the root mean square error (RMSE) is optained. In this study it will be used as a performance measurement.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}. \tag{2.64}$$

The MSE, and RMSE as well, alone does not give any clear picture of a performance for any single model, it is more appropriate when comparing two or more models fitting some actual data. As a measure for the single model it is more suited to use the *coefficient of determination* (R^2). It is defined as

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\Sigma_{i=1}^{K}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{K}(y_i - \bar{y})^2} \tag{2.65}$$

where SS_E is the sum of squared errors and SS_T is the total sum of squares of the response variable y. The coefficient of determination shows how well the model represents the data and is scaled from zero to one, one indicating 100% fit.

The R^2 is not totally accepted since it is not related to the number of parameters in the model and need therefore a little adjustment. The *adjusted* R^2 is a variation of the R^2 statistic compensates for the number of parameters in a regression model. The adjustment is denoted as

$$R^2_{adj} = R^2\frac{n-p}{n-1} \tag{2.66}$$

and it penalizes for adding terms to a model where n is number of observations and p is the number of parameters in the model.

Chapter 3

Varying-coefficient functions

3.1 Introduction

This class of models is a further generalization to models which are linear in some of the regressors, but the coefficients of the model are replaced by smooth, but otherwise unknown, functions of some other variables. The models are also called *varying-coefficient models* and are illustrated in detail in [9]. When all coefficients depend on the same variable, the model is denoted as *conditional parametric model*.

3.2 Locally weighted regression

Let y_i, for $i = 1, \ldots, n$, be the i-th measurement of the response and 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 (2.16):

$$y_i = g(x_i) + \epsilon_i, \tag{3.1}$$

where g is smooth function and the ϵ_i are random errors, i.i.d. and Gaussian. For the function assumed to be smooth allows points in certain neighborhood of x to estimate the response. This neighborhood is some fraction of the data closest to x where each point is weighted according to distance; increase in distance from x gives decrease in weight. The smoothing function is estimated by fitting a polynomial of the dependent variables to the response, $g(x) = P(x, x)$. For each fitting point, parameters of the polynomial ($\boldsymbol{\theta}$) need to be estimated, therefore locally-weighted least squares is considered:

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} w_i(\mathbf{x}) \left(y_i - P(\mathbf{x_i}, \mathbf{x})\right)^2 . \tag{3.2}$$

The local least squares estimate of g($\mathbf{x}$) is $\hat{g}(\mathbf{x}) = \hat{P}(\mathbf{x}, \mathbf{x})$. 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 locally weighted regression requires a weight function and a specified neighborhood size. To allocate the weights, $w_i(\mathbf{x}$, to the observations nowhere increasing weight function, W, is used. There are several weight functions which can be used and are few listed in Table **??**. In the case of spherical kernel the weight

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	$W(u) = \exp(-u^2/2)$

Table 3.1:

on observation i is determined by the Euclidean distance between $\mathbf{x_i}$ and $\mathbf{x}$, i.e.

$$w_i(\mathbf{x}) = W\left(\frac{\|\mathbf{x_i} - \mathbf{x}\|}{h(\mathbf{x})}\right). \tag{3.3}$$

The scalar $h(\mathbf{x})$ is called the bandwidth and is greater than zero.

The bandwidth is an indicator for the neighborhood size. If it is 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,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.

3.3 Conditional parametric 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 other variables. These kind of models are called *varying-coefficient models* [9], but when all coeffiecients depend on the same variable the model is referred to as *conditional parametric model*.

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.

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta}(\mathbf{u}_i) + e_i, \tag{3.4}$$

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 $\boldsymbol{\theta}(\cdot)$ in (3.4) are estimated at a number of destinct points, fitting 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 $\mathbf{u} = [u_1 \quad u_2]_T$, then $\mathbf{p}_2(\mathbf{u}) = [1 \quad u_1 \quad u_2 \quad u_1^2 \quad u_1 u_2 \quad u_2^2]$. Also let $\mathbf{x}_i = [x_{1,i} \cdots x_{p,s}]$. Then

$$\mathbf{z}_i^T = \left[x_{1,i}\mathbf{p}_{d(1)}^T(\mathbf{u}_i) \cdots x_{j,i}\mathbf{p}_{d(j)}^T(\mathbf{u}_i \cdots x_{p,i}\mathbf{p}_{d(p)}^T(\mathbf{u}_i \right] \tag{3.5}$$

and

$$\boldsymbol{\phi}_u^T = \left[\boldsymbol{\phi}_{u,1}^T \boldsymbol{\phi}_{u,j}^T \boldsymbol{\phi}_{u,p}^T \right], \tag{3.6}$$

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

$$y_i = \mathbf{z}_i^T \boldsymbol{\phi}_u + e_i \tag{3.7}$$

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

$$\hat{\boldsymbol{\phi}}(\mathbf{u}) = \arg\min_{\boldsymbol{\phi}_u} \sum_{i=1}^{N} w_u(\mathbf{u}_i) \left(y_i - \mathbf{z}_i^T \boldsymbol{\phi}_u \right)^2, \tag{3.8}$$

for which a unique closed form solution exists provided the matrix with rows $\mathbf{z}_i^T$ corresponding to non-zero weights has full rank. The weights are the same as illustrated in section 3.2 above. The elements of $\boldsymbol{\theta}(\mathbf{u})$ are estimated by

$$\hat{\theta}_j(\mathbf{u}) = \mathbf{p}_{d(j)}^T(\mathbf{u}) \hat{\phi}_j(\mathbf{u}) \tag{3.9}$$

where $j = 1, \ldots, p$ and $\hat{\phi}_j(\mathbf{u})$ is the weighted least squares estimates of $\phi_{u,j}$.

When $\mathbf{z}_j = 1$ for all j this method is identical to the locally weighted regression described above.

3.4 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

$$\hat{\boldsymbol{\phi}}_t = \arg\min_{\boldsymbol{\phi}} \sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s) \left(y_s - \mathbf{z}_s^T \boldsymbol{\phi}\right)^2 \tag{3.10}$$

where $w_u(\mathbf{u}_s)$ is a weight on observation s depending on the fitting point $\mathbf{u}$ and $\mathbf{u}_s$.

The adaptive estimates in (3.10) can be found recursively as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) + w_u(\mathbf{u}_t)\mathbf{R}_{u,t}^{-1}\mathbf{z}_t \left[y_t - \mathbf{z}_t^T \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})\right] \tag{3.11}$$

and

$$\mathbf{R}_{u,t} = \lambda \mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T. \tag{3.12}$$

Note that $\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is a predictor of y_t locally with respect to $\mathbf{u}$ and for this reason it is used in (3.11). To predict y_t a predictor like $\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is appropriate.

The method of adaptation for local estimation is not applied in this presentation. For more details on time-varying coefficient functions see [12] or [8].

3.5 The LFLM library in S-PLUS

In S-PLUS the loess function can be used for estimation in some simple conditional parametric models. In one of the papers in [14] H.A. Nielsen describes a software implementation he developed which deals with conditional parametric models. This software package runs under S-PLUS and has the advantage of having no restriction on the global linear model, which is formulated along with the local model. In loess the originating model need to be straight line or a linear hyperplane.

The name of this S-PLUS library is LFLM (Locally weighted Fitting of Linear Models) and can be downloaded from author's homepage[1]. LFLM can though be a bit complex when simple local fit is needed, thus both functions were used in analysis of this thesis.

[1] http://www2.imm.dtu.dk/ han/software.html

Chapter 4

Data

4.1 Introduction

The data used in the analysis are power predictions from WPPT (Wind Power Predictions Tool) for the production at Klim wind farm, which is located at the northwest coast of Jutland. Running meteorological (MET) forecasts from various MET forecast systems into WPPT, results in different wind power predictions for the wind farm. These predictions are combined for more accurate forecast.

WPPT is a system for forecasting the wind power up to 48 hours ahead depending on the horizon of the MET forecasts, It is able to forecast for wind power production in relatively large regions and for individual wind farms. WPPT uses the wind speed and wind direction from the MET forecasts as inputs. To read more about WPPT see [13].

The data is twofold. Section 4.2 describes the forecasts from WPPT, whilst section 4.3 illustrates the meteorological forecasts which are used to analyse the weights in the combined forecast.

4.2 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 where 35 600kW V44 rotors are generated and the predictions from WPPT range from 0 to 21000kW. Thus, this is the range available for the (absolute) prediction errors as well.

The aim is to model a combined forecast, denoted as **comb.fc**, with the forecasts from WPPT as the explanatory variables, which are represented as

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

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

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

The forecasted power is given every hour for all weather systems and is based on forecasts at time point 00Z. From midnight the power predictions are given with a 24 hour horizon. There are 7272 data points in the data set which span the period from February 2nd 2003 to December 2nd 2003.

The three different forecasts are based on different meteorological data and since all predicting for the same event, they are very correlated. This correlation can be identified from Figure 4.1 which shows the pairwise scatterplot of the predicted power from the three different WPPT forecasts. The correlation for all three forecasting systems is approximately 0.85.

The 24 prediction horizons are also a variable in the analysis and are denoted with **horizon**. The data set is divided w.r.t. horizons and behaviour within each horizon investigated. The issue of missing data can influence the horizon variable if the difference in number of observed values in each horizon is large. Figure 4.2 show how the individual forecasts are divided to the horizons. From the figure it is observed that MM5 (right) has most of missing data, while DWD (left) has most valid data. The valid data in HIRLAM (middel) is little less then for DWD, but there are more missing values for the same horizons. Non of the forecasts have big difference in horizons such that it would influence the investigation.

Figure 4.3 shows the time series for all three competing forecasts. As men-

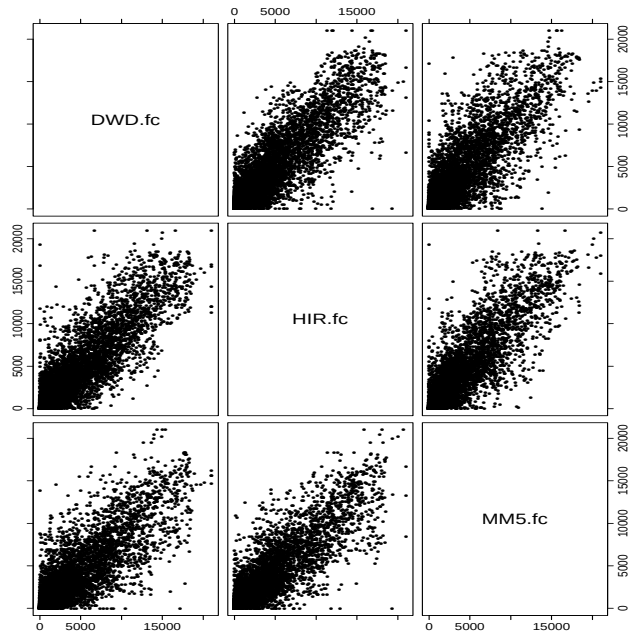


Figure 4.1: Pairwise scatterplot of the competing forecasts.

tioned before they all predict for the same event and therefore they all look similar. But there are departures in the processes, specially when MM5 is compared with the other two. When high production is expected MM5 appears to

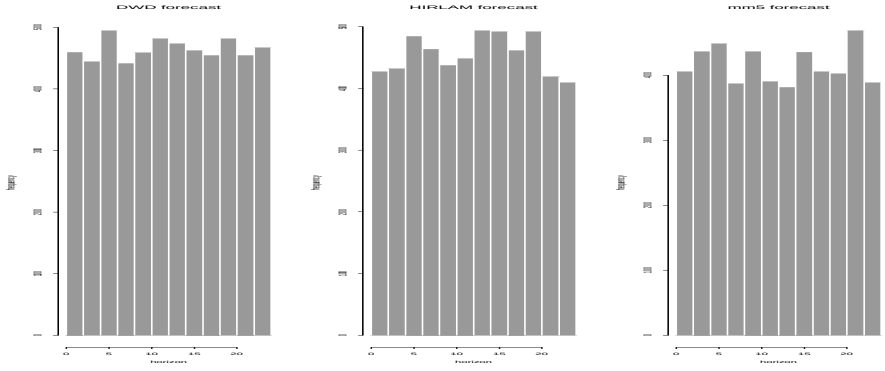


Figure 4.2: Histograms of number of valid forecasts within the individual forecasts; DWD (left), HIRLAM (middle) and MM5 (right)

forecast lower power production. The time series also shows how frequently the MM5 forecast indicates missing values, f.ex. end of Mars, middle of June and twice in late November.

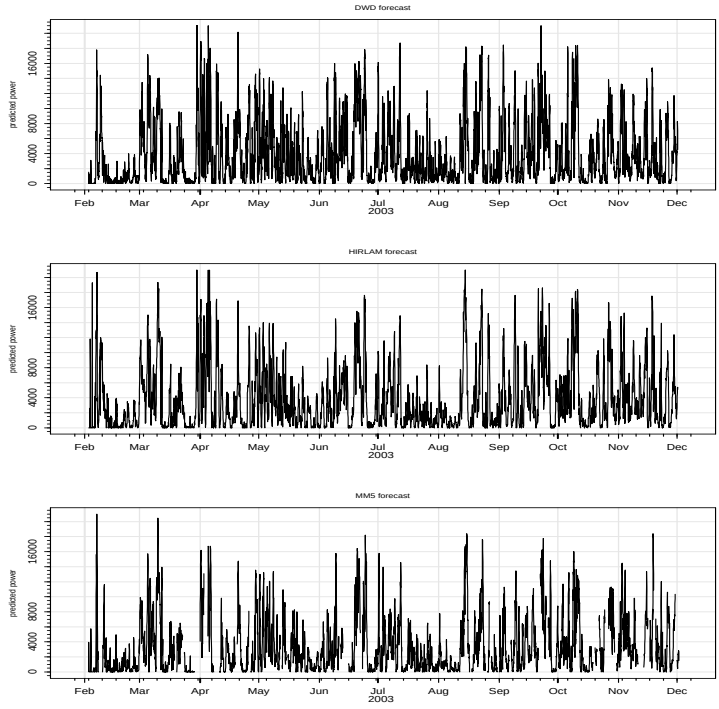


Figure 4.3: Timeseries plots of the individual forecasts

4.3 Meteorological data

The data set consist of meteorological data from DMI's meteorological forecasting system DMI-HIRLAM. The meteorological forecasts are only available in 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 location of Klim and the grid points around is showed in Figure 4.4.

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 ex-

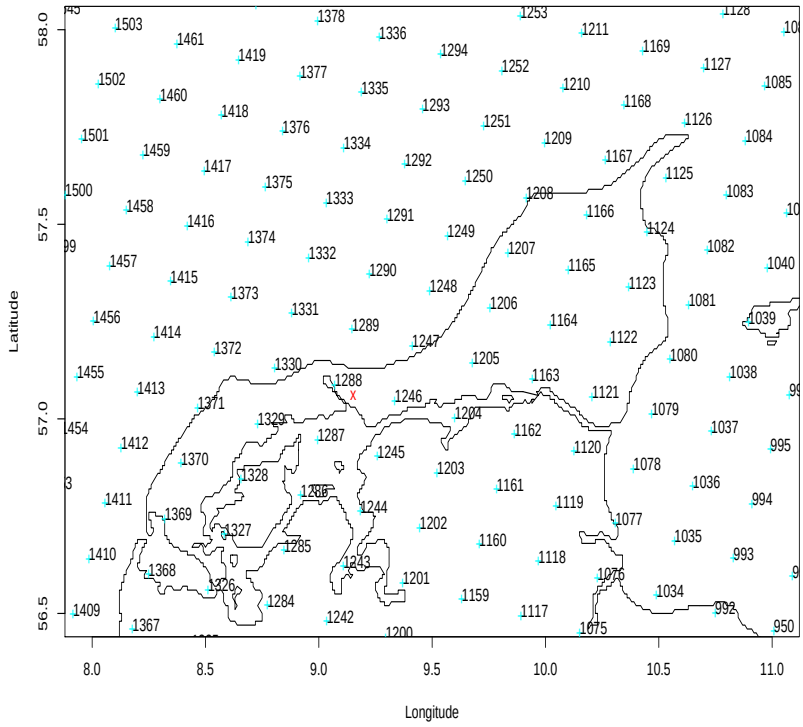


Figure 4.4: The grid points where meteorological forecasts are available. Klim wind power plant is marked with red.

planatory variables which are found in a set of the meteorological forecasts. In the data set most of the meteorological forecasts are available every hour. Those who are only accessible from the archive data files, arrive every 3rd hour. Hourly predicted variables are the following:

ws10m: Forecasted wind speed at 10 meters above ground level (m/s).

wd10m: Forecasted wind direction at 10 meters above ground level (degrees).

rad: Forecasted radiation (W/m^2)

fv: Forecasted friction velocity (m/s).

ad: Forecasted air density (g/m^2).

and from the archive, every 3rd hour the variables in the data set are:

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

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

tkeL··: Forecasted turbulent kinetic energy in model level ··, the levels in the data set are 31, 38, 39 and 40 (Wm^2/s^2).

The model levels are different pressure levels of the atmosphere. The numbers position the levels, with increasing number there is a decrease in height above ground. All MET forecasts have 1 hour resolution except the archive data which has 3 hour resolution. By linear interpolation 1 hour prediction is generated for the archive forecasts.

In the analysis of the MET forecasts not all variables are used due to similarities, only those depicted in the pairwise scatterplot in Figure 4.8 are of interest. 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 tubulent kinetic energy are all alike and thus only on is applied, at level 38. In Appendix A the similarity within these variables is illustrated with pairwise scatterplots. Also in Appendix A the complete list of the WPPT forecasts and the MET forecasts is displayed.

The MET forecasts from DMI-HIRLAM are given with 48 hours horizon but in this presentation only the 24 hours horizon is applied, forecasted at 00Z and 24 hours ahead.

Figure 4.5 illustrate histograms for some of the MET forecasts. Both wind speed variables, (**ws10m** and **wsL31**), along with friction velocity (**fv**) appears to be normally distributed but skewed towards the origin. The pairwise scatterplot in Figure 4.8 shows **ws10m** and **fv** being very correlated. The friction velocity can be thought of as a wind speed and therefore either **ws10m** or **fv** is chosen as an explanatory variable. The wind speed is chosen and friction velocity therefore omitted in the analysis. The air density (**ad**) appears to be normally distributed but the radiation (**rad**) looks like being in two stages, less and more than 100. The turbulent kinetic energy (**tkeL38**) is more densed at lower values and then reduces rapidly, exponential distribution.

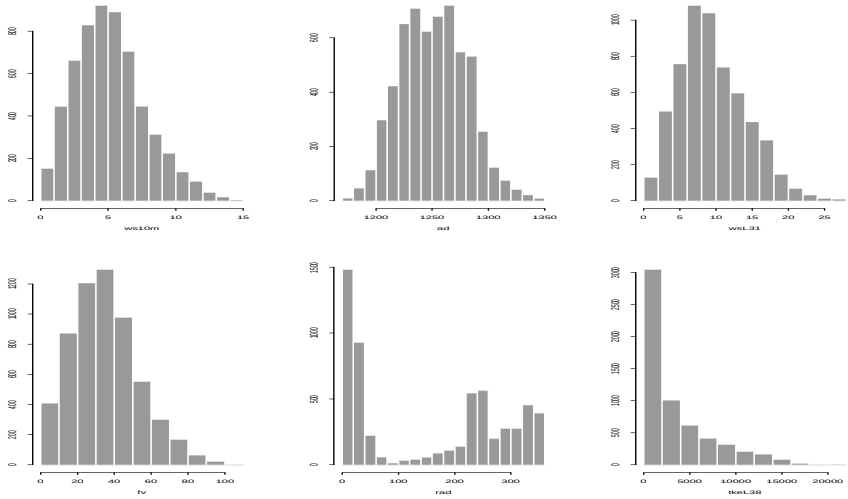


Figure 4.5: Histogram of some of the explanatory variables.

When analyzing the MET data it might be convenient to consider some of the variables as factors. The directional variables can be catagorized in four main directions, north, south, east and west. Zero degrees is straight north and than the degrees increase clockwise. This will divide the 360 degrees circle as $0 \leq N \leq 45$, $45 < E \leq 135$, $135 < S \leq 225$, $225 < V \leq 315$ and then close the circle with $315 < N \leq 360$. The letters indicate corresponding direction. Figure 4.6 illustrates how the data is characterized.

Figure 4.7 shows the frequency of the four factors defined for wind direction. Since the wind power plant is located near the west coast by the Northsea, the most frequent wind direction is west both for directional variables at sea level (**wd10m**) and up in the atmosphere (**wdL31**). The only difference between the two variables is the wind blowing on the boundary between E and N, wind coming from north in higher hemispheres tends to rotate few degrees so it blows from east closer to the ground.

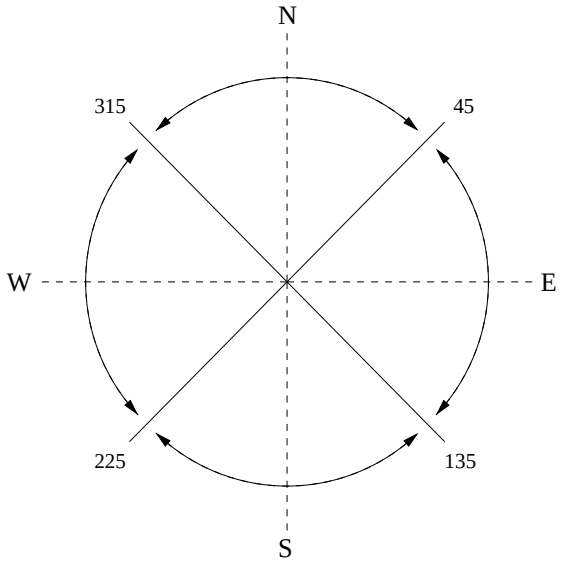


Figure 4.6: Illustrates how the degrees in wind direction is divided to four components (N,E,S,W)

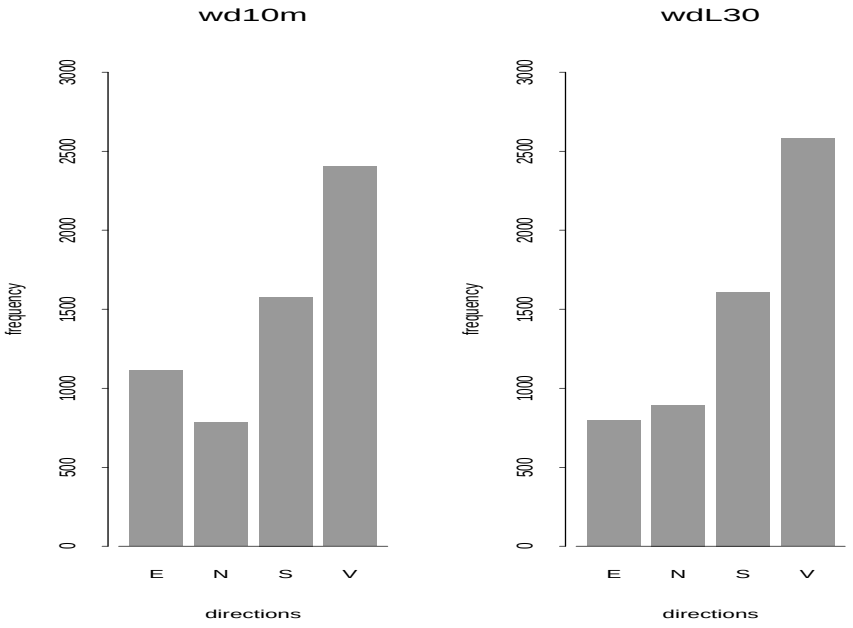


Figure 4.7: Histogram of the direction variables as factors.

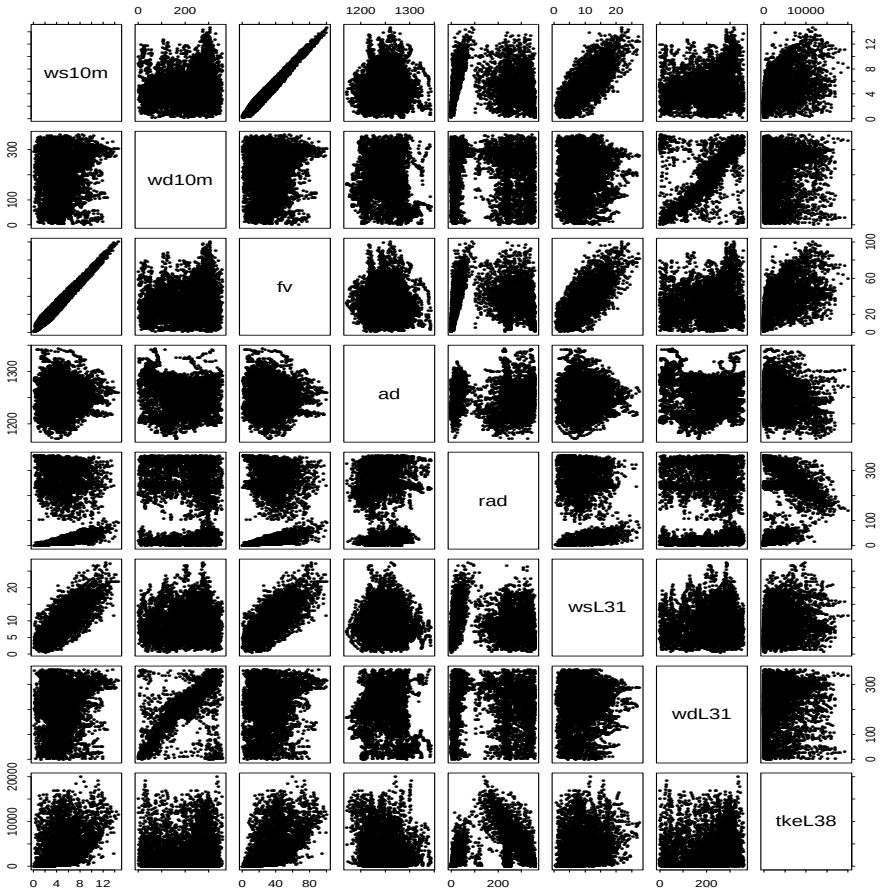


Figure 4.8: Pairwise scatterplot of the meteorological data.

Chapter 5

Combining WPPT forecasts

5.1 Introduction

The chapter focus on combining the WPPT forecasts, listed in section 4.3 with the methods introduced in chapter 2. Wind power prediction is a tool which works on-line in an environment with minimum user interference for long period of time. Time evolution in the dynamic properties of the predictor should be allowed which indicates that the estimated weights of the combined forecast should be time-varying.

The recursive methods applied in this chapter are RLS method and the adaptation of the optimal method. The performance of the simple average method is also applied for comparison.

In section 5.2 the individual forecasts are inspected. A small discussion about when constant can be added to the regression model of combining and some argument for assuming the weights summing to one is stored in 5.3. In 5.4 the static model for combining is considered as a reference and in section 5.5 the on-line combination is analyzed after the forgetting factor has been chosen.

5.2 Individual forecasts

By combining forecasts a prediction is wanted which perform better or at least equal to the individual forecasts. But do the individual forecasts imply any information about which two is optimal to combine? In this section 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 most adequate forecast is the forecast closest to the observed values, thus the forecast errors are distance measures of the accuracy of a forecast. RMSE, defined in section 2.10, interprets this distance for the whole data set and in Figure 5.1 the RMSE for the forecasts are depict for all horizons. As expected,

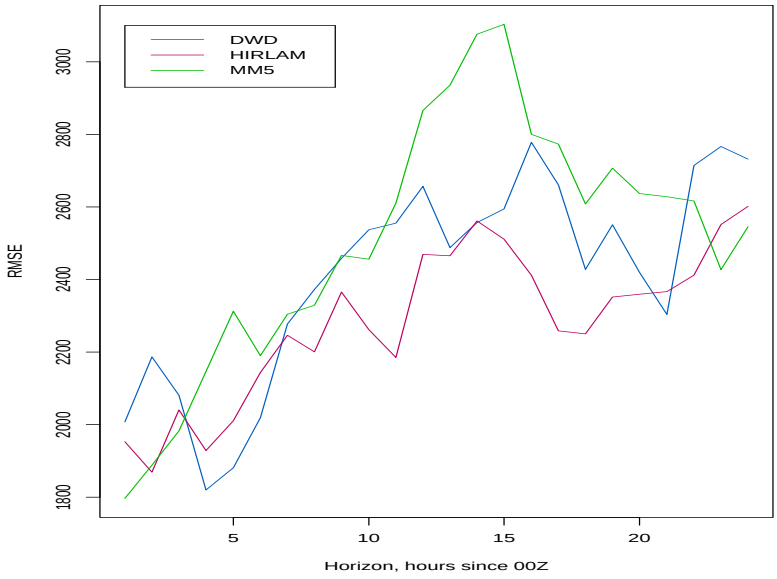


Figure 5.1: RMS for the individual forecast errors

RMS of the errors are low in the first 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 then HIRLAM. MM5 forecast is the least rigorous prediction of all three. As Figure 5.1 indicates, MM5 is really bad presentative for forecasting in horizon 11 to 15. It would be interesting to see which two competing forecasts is best to combine, especially

between horizons 11 to 20 since the difference in RMSE within forecasts appear to be the greatest in these horizons.

From Figure 5.1 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. The correlation is, however, a good representative to compare inner structure of two processes. Therefore it would be interesting to see if the correlation of the prediction errors can give any knowledge about the best combination. The correlation between every pair of forecasts, for all horizons, is shown in Figure 5.2. The figure indicates that all sets of forecasts correlates quite similar,

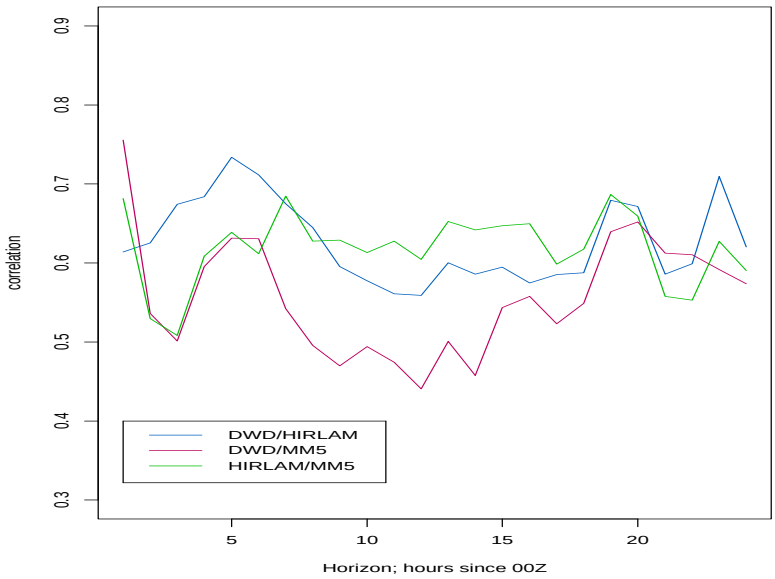


Figure 5.2: Correlation between individual forecasts

around 0.6. It is though worth to notice that the HIRLAM forecast correlated with the other two, variates less then the correlation between DWD and MM5. The correlation from horizon 6 to 15 is rather lower between DWD and MM5 than other correlations. Overall non of the combinations have high correlation within forecast errors.

5.3 Restriction and constant

In chapter 2 a restriction on the weights is introduced when regression model is used to aggregate forecasts. The restriction can be applied to the model in many ways, e.g. with equality restricted least squares method as in [7] where the weights are also restricted to be non-negative. The problem then become an optimality problem where Kuhn-Tucker constraints are applied. The modelling for the restriction in this presentation is more straight forward method simply by adding the constraint into the combination model in (2.2); with K individual forecasts and the restriction $\Sigma_{i=1}^{K} w = 1$ the model becomes

$$y = w_0 + w_1\hat{y}_1 + \cdots + w_{K-1}\hat{y}_{K-1} + \left(1 - \sum_{i=1}^{K-1} w_i\right)\hat{y}_K + e \tag{5.1}$$

$$= w_0 + w_1\left(\hat{y}_1 - \hat{y}_K\right) + \cdots w_{K-1}\left(\hat{y}_{K-1} - \hat{y}_K\right) + \hat{y}_K + e \tag{5.2}$$

and by subtracing with the K-th forecast

$$y - \hat{y}_K = w_0 + w_1\left(\hat{y}_1 - \hat{y}_K\right) + \cdots + w_{K-1}\left(\hat{y}_{K-1} - \hat{y}_K\right) + e \tag{5.3}$$

$$\tilde{y} = w_0 + \sum_{i=1}^{K-1} w_i\tilde{\hat{y}}_i + e. \tag{5.4}$$

From the model evalution above neither the prediction error e nor the constant w_0, if included, are affected by the modification. Thus not only the weights are properly estimated, but also the restricted model in (5.4) 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. If the ingredient forecasts are "good" forecasts this is not of concern, but the constant is included to detect any abnormal behavior of the combination.

In [11] there is a discussion on difference between unrestricted regression method and adding the restriction to the combination. It is noticed that when the regression method is restricted and is without an intercept, it can be viewed as the optimal method. This is likely to give worse accuracy for the combining than for the unrestricted regression. But there are few issues related to the unconditioned regression based method; stationarity of the forecasted variable, autocorrelation and multicollinearity might be of concern when individual forecasts are correlated which is often the case when combining.

In [6] Granger et al. concluded that a reasonable approach to combining is to include a constant in the regression and restrict the weights on the forecasts to add up to one. They argued that if the individual forecasts are biased, restricted regression including constant should outperform the optimal method.

The constant debiases the individual forecasts in the combination. de Menezes et. al. [11] claimed this not totally correct since the constant will only debias for location bias, but not scale bias. To see the bias in the individual forecasts the mean square errors of the predictions is viewed as

$$\begin{aligned} MSE[\hat{y}] &= E[(\hat{y}-y)^2] \\ &= E\left[(\hat{y}-E[\hat{y}]+E[\hat{y}]-y)^2\right] \\ &= E\left[(\hat{y}-E[\hat{y}])^2\right]+(E[\hat{y}]-y)^2 \\ &= V[\hat{y}]+b^2. \end{aligned} \tag{5.5}$$

where $\hat{y}$ is the individual forecast and y is the observations. From (5.5) it can be seen that the bias is the difference between the forecast error variance and the forecast MSE.

Checking for bias is one way of investigating if constant is needed in regression, another way is to include the constant and see if it is signifcant in the parameter estimation.

In this presentation the consideration of a constant term in the model is controlled by the mean forecast error of the individual forecasts. Figure 5.3 displays the mean error for the competing predictions for all 24 prediction horizons. Any deviation for the mean forecast errors from zero indicates that constant is needed, but for an adequate forecast the prediction error is close to nil. What Figure 5.3 illustrates is that for any individual forecasts to predict for the wind power at Klim, an intercept is involved to explain the departure of the mean prediction error. This is valid for almost all prediction horizons for all possible combinations. For horizon 7 through 16 the DWD and MM5 forecasts appears to have similar magnitudes but in opposite directions. Combining these two might reduce the importance of intercept in the combination. Combining DWD and HIRLAM could present more stability of the intersept through the prediction horizons since the forecasts oscillate in opposite phases.

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.

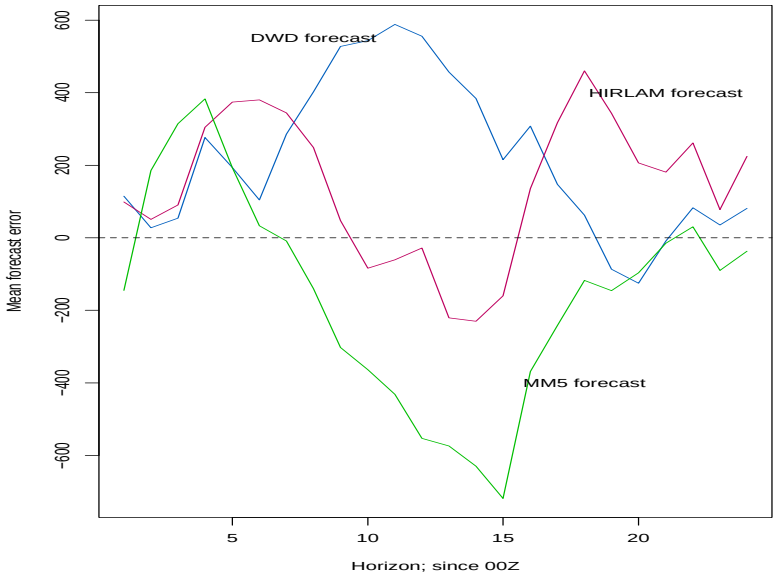


Figure 5.3: Mean forecast error for the individual forecasts for each prediction horizon

5.4 Offline combination for wind power forecasts

In this section an off-line estimation of the parameters (weights) in the regression model in (2.12) is considered. The coefficients for the model are estimated for the whole period from February to December. The normal procedure when estimating off-line is to divide the data set in two subsets where one part is used to estimate the parameters in a regression model and the second part to validate the model.

Considering the regression model in (5.4) to estimate the weights allows the parameters to be estimated over the entire data set. The validation takes place by applying the estimated parameters from (5.4) to the regression model in (2.12). For each horizon a restricted regression model is considered to estimate the weights by using the least squares method where the quadratic loss function, defined in section 2.6.1, is denoted as

$$S(\mathbf{w}) = \mathbf{e^T e} = (\tilde{y} - \tilde{\hat{y}}\mathbf{w})^T(\tilde{y} - \tilde{\hat{y}}\mathbf{w}) \tag{5.6}$$

with $\tilde{y}$ and $\tilde{\hat{y}}$ from (5.4). The estimated weights are then observed as

$$\tilde{\hat{\mathbf{w}}} = (\tilde{\hat{\mathbf{y}}}^{\mathbf{T}}\tilde{\hat{\mathbf{y}}})^{-\mathbf{1}}\tilde{\hat{\mathbf{y}}}^{\mathbf{T}}\mathbf{y}, \tag{5.7}$$

but since estimation from (5.7) only evaluate weights for the first $K-1$ forecasts, the Kth weight is considered by the restriction. The weights are then observed as

$$\hat{\mathbf{w}} = \begin{bmatrix} \tilde{\hat{\mathbf{w}}} \\ 1 - \sum_{i=1}^{K-1} w_i \end{bmatrix}. \tag{5.8}$$

and are used in the linear regression model to predict for the power production.

Estimated weights for the combined forecasts are displayed in Figure 5.4 evolving with prediction horizon. Each panel shows the quantitive weights in a combination. The panels appears to be reflecting the accuracy for the individual forecasts dipicted in Figure 5.1 when two forecasts are combined. Compositing the two best performing forecasts, DWD and HIRLAM, give weights fluctating around the equal weight at 0.5. By investgating panels two and three, combinations including the MM5 forecast, the pattern from Figure 5.1 is quite clear. For the horizons where the MM5 prediction is performing poorly, corresponding weights in the combinations is decreased down to 0.2. The same appears in the bottom panel where weights are displayed when conpounding all three forecasts. Again the bad performance of the MM5 forecast give significantly lower weight than the other ingredient forecasts. A weight of individual forecast in combination is reflected by the performance of the individual forecast compared to other constituent forecasts in the combination.

Figure 5.5 display the magnitude of a intercept, estimated for the combinations at each prediction horizon. It is noticed that most of the horizons have negative constant. It is only the HIRLAM/MM5 forecast which give great amplitudes on the positive side for horizons 13 to 16. Negative intercept means that the corresponding aggregation is over-estimated. Figure 5.5 indicates that the intercept can not be omitted.

5.4.1 Performance for constant weights

When the linear model for combining is considered the parameters are quantified as Figures 5.4 and 5.5 show for weights and an intercept respectively. In the case of combining DWD and HIRLAM the weights have similar behaviour 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 the two best performing individual forecasts, the DWD/HIRLAM combination appears to have the lowest coefficient

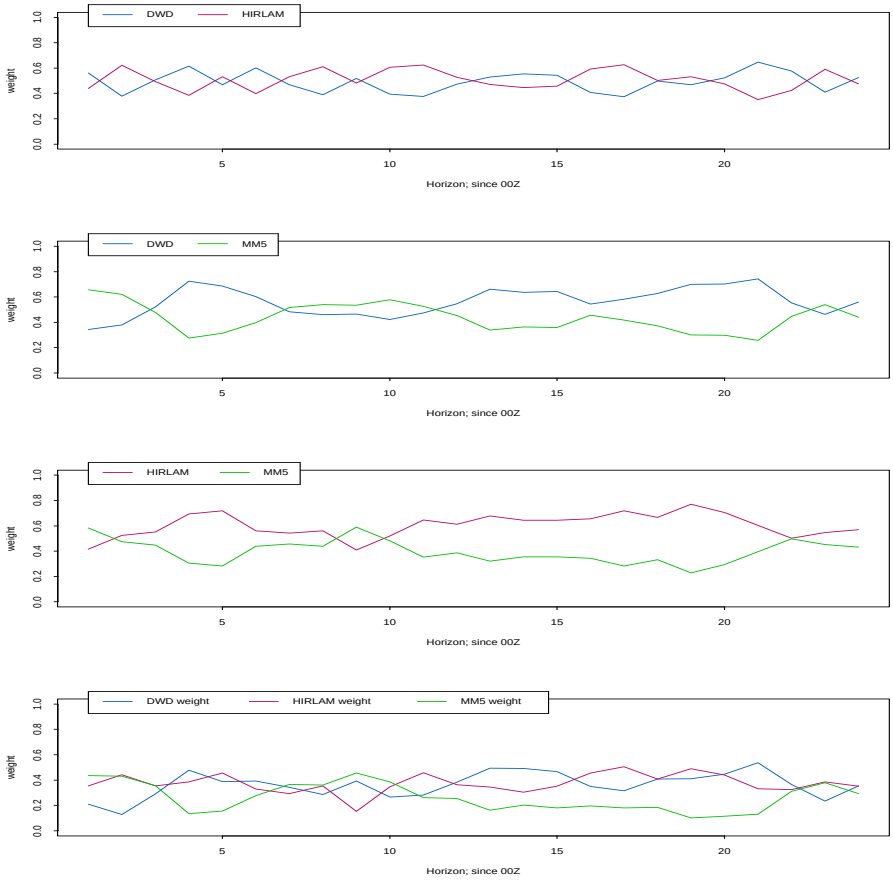


Figure 5.4: Size of the weights in a combined forecast. Each panel shows how the weights change with prediction horizons in a combination. The legends in the panels indicate what individual forecasts are combined.

of determination as indicated in Table 5.1. For horizons listed in the table, is DWD/HIRLAM 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 the 6th horizon. 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 extra forecasts.

Figure 5.6 shows RMSE for all off-line combinations in all 24 prediction horizons.

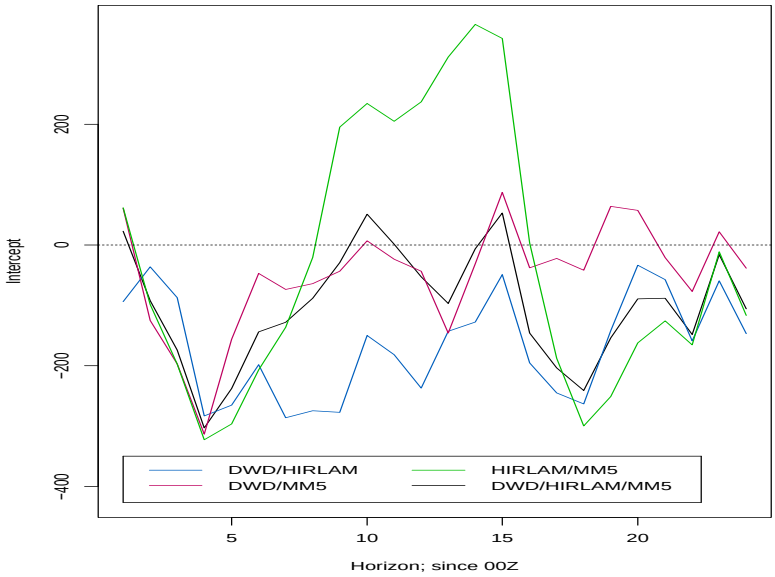


Figure 5.5: Estimated intercept in each horizon for the combined forecasts. The estimations can not be neglected.

It confirms the best performance of the HIRLAM/MM5 forecasts in the early horizons til DWD/HIRLAM become the best performing prediction. However, what Figure 5.6 shows is that after horizon 15, DWD/HIRLAM outperforms the other two.

For the entire prediction horizon, the combination of all three competing fore-

Combination	Prediction horizon						
	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

Table 5.1: 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

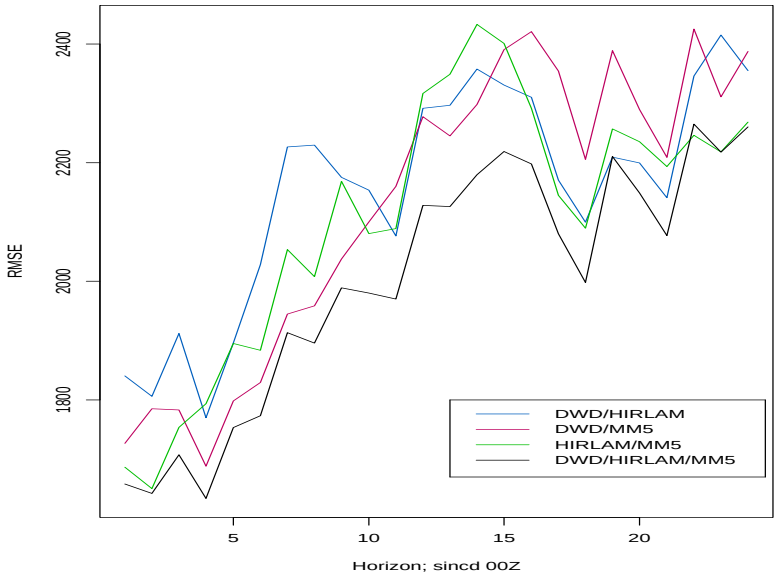


Figure 5.6: In-sample RMSE for combined forecasts over the prediction horizons in an offline estimation.

casts outperform the compositing of two. This is not surprising since informations from all three are gathered to improve the accuracy. It is though noticed from Figure 5.6 that for the first few horizons and the few last, the performance of DWD/HIRLAM/MM5 is only a little better than the best performing synthesis of two forecasts.

5.5 Online combination for wind power forecasts

When combinining wind power forecasts where new data is added to the data frequently, it is appropriate to estimate the weights adaptively. Estimating the parameters with an off-line approach is not an adequate procedure and need to be extended to allow time-variation.

The methods described in section 2.8 are applied to estimate the weights recursively, but addition is needed for the RLS method in (2.58) and (2.59) since the weights are restricted to sum to unity. The model which is estimated re-

cursively is the restricted linear model (5.4) where the Kth weight at time t is calculated by $\Sigma_{i=1}^{K} w_{i,t} = 1$. The restricted RLS algorithm is

$$\tilde{\hat{\mathbf{w}}}_t = \tilde{\hat{\mathbf{w}}}_{t-1} + \mathbf{P}_t \tilde{\hat{\mathbf{y}}}_t \left[y_t - \tilde{\hat{\mathbf{y}}}_t^T \tilde{\hat{\mathbf{w}}}_{t-1} \right] \tag{5.9}$$

$$\mathbf{P}_t = \frac{1}{\lambda} \left[\mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1} \tilde{\hat{\mathbf{y}}}_t \tilde{\hat{\mathbf{y}}}_t^T \mathbf{P}_{t-1}}{\lambda + \tilde{\hat{\mathbf{y}}}_t^T \mathbf{P}_{t-1} \tilde{\hat{\mathbf{y}}}_t} \right] \tag{5.10}$$

where $\tilde{\hat{\mathbf{y}}}$ is a vector of forecast differences as defined in (5.4), $\tilde{\hat{\mathbf{w}}}$ is a vector of the first *K-1* forecast weights (and intercept if included) and λ is the forgetting factor. The combined forecast at time t is then estimated by

$$\hat{y}_{c,t} = \tilde{\hat{\mathbf{w}}}_t^T \tilde{\hat{\mathbf{y}}}_t + \hat{y}_{K,t}. \tag{5.11}$$

Choosing a forgetting factor for the adaptation is important since it represents influence of past data. It is desided to keep it constant for all prediction horizons but further inspection in needed.

5.5.1 Selecting the forgetting factor

The choice of appropriate forgetting factor is a key feature of adaptation since it has a substantial effect on the efficiency of the predictions. Two of the combination methods, regression and the optimal method, generate forgetting factor in adaptive estimation.

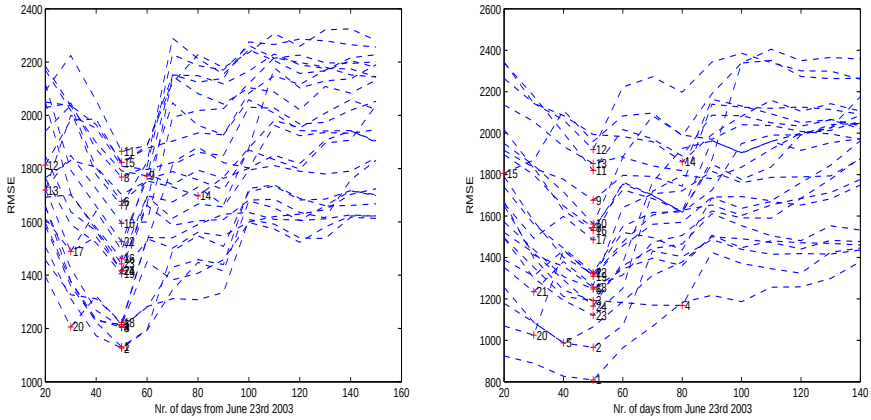
When combining the forecasts from WPPT two issues are of concern. First is the issue of choosing a forgetting factor due to the horizons. Each horizon has time-varying weights based on λ, evaluated within the horizon. Second is the choice of forgetting factor for different combinations.

The procedure for selecting the forgetting factor is to use the first part of the data set for the choise. Then use the observed λ for the whole data set. The appearance of missing values in the data is a factor in the selection. Therefore it is more sufficient to use the longest period without missing data for evaluation. In the data set there are forecasts and observations which are missing. The length of periods with non-missing values differ with respect to what individual forecasts are combined. The MM5 has the most missing data and when used in combining, the period of non-missing values in the composite forecast is shorter than other possible combinations of competing predictions.

Combined DWD and HIRLAM forecast has non-missing values from June 16th and 133 or 151 days ahead depending on the prediction horizon. Including MM5

forecast in combination reduces the non-missing interval substantially. It start a week later, 23rd of June, and only stays for 56 days for all horizons. Evaluating λ for more the 56 days might be influenced by the missing values.

The part of the data set used to estimate λ is initiated on the 23rd of June and spans the following 150 days. This part is also used to find forgetting for the combination of DWD and HIRLAM though the data set is valid from the 16th of June. The plots in Figure 5.7 show two different methods of combining

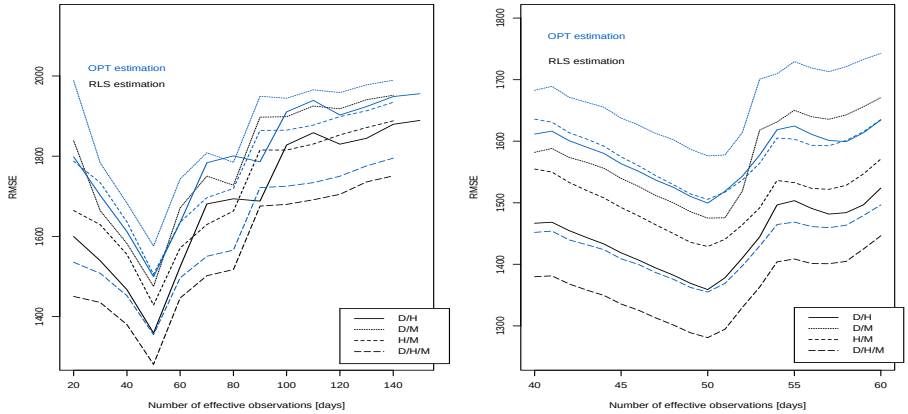


(a) D/H with OPT method (b) D/M with RLS method

Figure 5.7: Minimum RMSE used to estimated the number of effective days. For every prediction horizon (dashed lines) the minimum RMSE is located with a red mark. Two possible combining predictions are depicted.

where the dashed lines are RMSE for every horizon in the combination method. Each mark (red), along with a horizon number in the graphs, indicates the location of the minimum RMSE for the horizons. The figures show the minimum RMSE are placed around 50 days from the initial day for most of the horizons. It can therefore be assumed that instead of use a forgetting factor for each horizon, one forgetting is generated to represent all horizon.

In Figure 5.8(a) the average RMSE for each day is displayed for all possible combinations, for the 150 days used to estimate λ. Two methods are displayed and show parallel behavior, but that is due to the intercept in the formulation of the RLS method. What Figure 5.8(a) illustrates is that all combinations appear to have a minimum RMSE close to 50 days. This should not be surprising since most of the minimum values for the horizons are close to 50 days as Figure



(a) Minimum RMSE over 150 days (b) Minimum RMSE over days 40 to 60

Figure 5.8: Minimum RMSE used to estimated the number of effective days. Average RMSE each day plotted for different methods and combinations. In (a) the estimation period is 150 days with 10 days resolution, but in (b) the eatimation period is narrowed around 50 days with resolution of 1 day.

5.7 displays. These estimations have resolution of 10 days and to find precise forgetting factor for the combined forecasts, the interval between 40 days and 60 days is inspected further. Figure 5.8(b) shows this interval of evaluation and illustrates that only the DWD/MM5 forecast has the minimum RMSE at day 51, other combinations have 50. Figure 5.8(b) shows also that the difference between RMSE for 50 days and 51 days in the DWD/MM5 aggregation is small. Thus, it can be concluded from Figure 5.8(b) that the minimum distance between the prediction and the oberved values appears when past 50 daily observations are used to estimate the weights. These 50 days give a forgetting factor of $\lambda = 0.98$. This forgetting is the objective for all methods and possible combinations.

5.5.2 Tracking time-varying weights

When the weights are estimated for all time points in the data set, traces of the weightsa are formed over the whole data set. The traces are used to observe the behaviour of the weights over time and check for stability in the parameters. This can then be compared to the estimated weights for the static model.

The time-varying weights for the combined forecasts are plotted in Appendix B but the plots span only 245 days for each horizon (the whole data set spans 303 days per horizon). This is because the first two months of the data are used to adjust the forgetting factor.

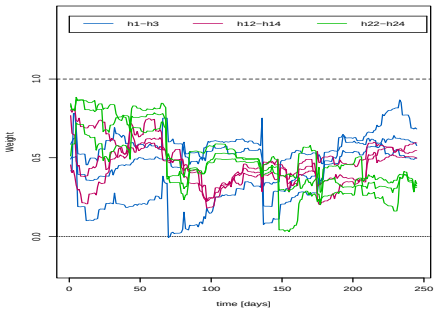
For the DWD weights (Figures B.1, B.2 and B.4) the behavior is quite similar. In the first horizons the weights are either unstable or close to zero in the beginning. With time the effects of the DWD forecast increases and stablizes around 0.5. For some horizons in DWD/HIRLAM is the weight a little higher. With prediction horizons 9 to 16 the influence of DWD is quite stable over time (middle panels) but for horizons over 16 the weights start of with great effects but decrease with time.

The HIRLAM weights (Figures B.3 and B.5) show opposite evolution with DWD. In the DWD/HIRLAM combination this is obvious since the sum of the weights is equal to one so the traces for the two forecast weights are mirrored around the average weights (0.5). When HIRLAM is combined with MM5 the weights are very stable for the first 8 and last 8 horizons. It is only for the intermediate horizons where the effects of HIRLAM starts of at low weights but over time it progresses. Very similar bahaviour appears for HIRLAM when combined with both DWD and MM5, as it was combined only with DWD. The effects of HIRLAM is more important in the beginning of the first 8 horizons but then decreases with time. The improvement for the last horizons is clear, but prediction horizons 9 to 16 show quite resembling behaviour as the HIRLAM does in the DWD/HIRLAM combination.

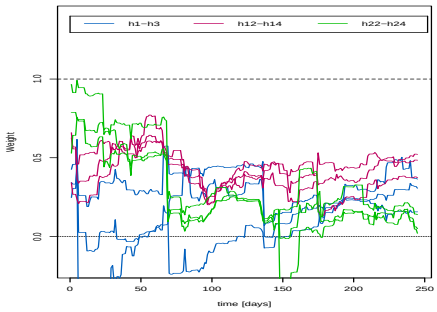
When all three forecasts are combined the MM5 weight eventually attain the same magnitude for nearly all horizons, approximately 0.3.

It is noticed from the all the figures that the tracking for the first 4 horizons are very unstable. The time-steps are varying rapidly and it appears that the weights are more sensitive at lower horizons. Despite that, the weights in all combinations accomplish some stability over the last 20 days.

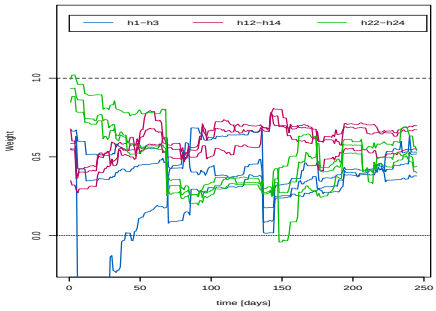
Figure 5.9 shows the first, intermediate and last three time-varying weights for each forecast weight from Appendix B. The DWD weights are displayed with various starting weights, but with time the coefficients become quite stable where all horizons have similar weights. It is only when DWD is combined with HIRLAM that the weights differ, DWD forecast have more effects on lower horizons. All possible combinations with HIRLAM show the same structure where HIRLAM has small influence within corresponding combined forecast for lower horizons, but the influence increases when the prediction horizon enlarge.



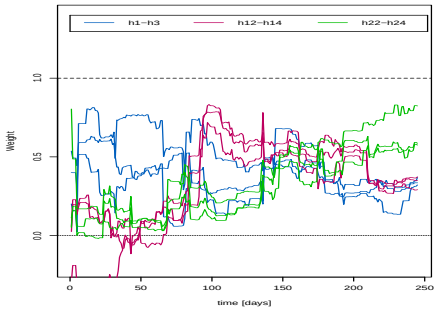
(a) DWD weights in D/H



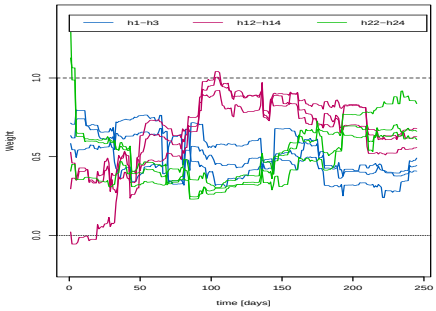
(b) DWD weights in D/H/M



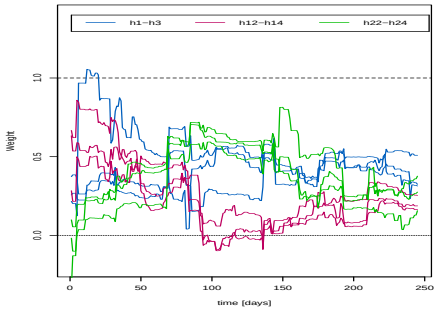
(c) DWD weights in D/M



(d) HIRLAM weights in D/H/M



(e) HIRLAM weights in H/M



(f) MM5 weights in D/H/M

Figure 5.9: 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.

The intercept for the combined forecasts are depicted in Figure 5.10. Each panel shows the time evolution for 24 intercepts in RLS estimation of a com-

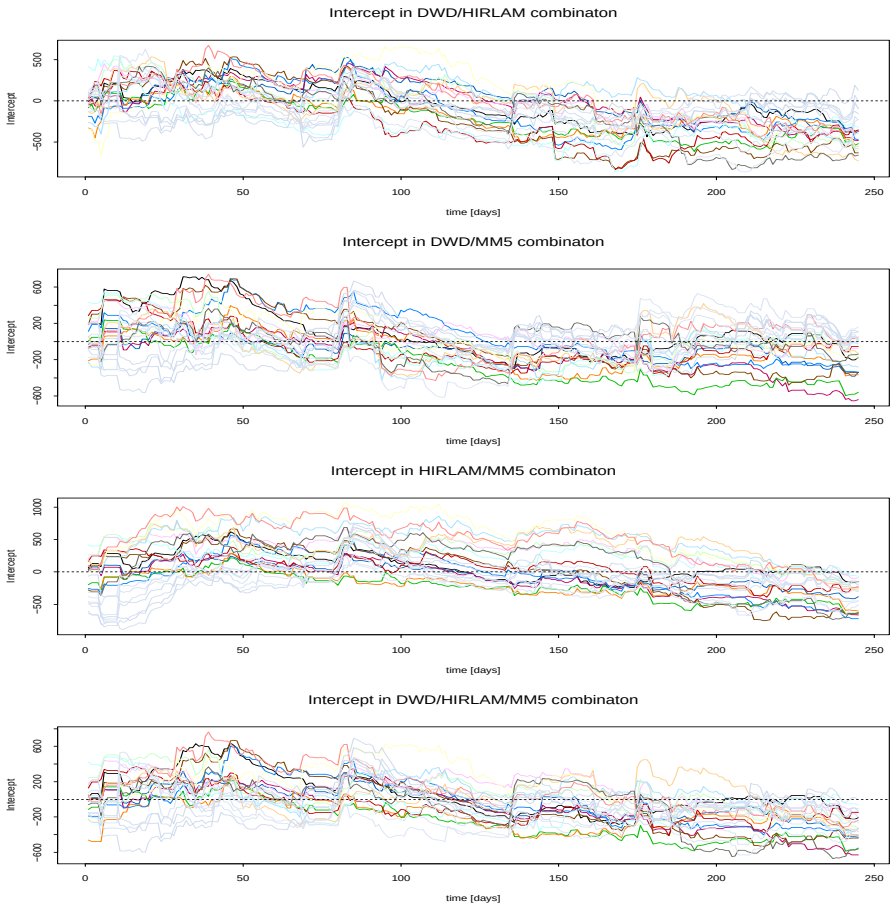
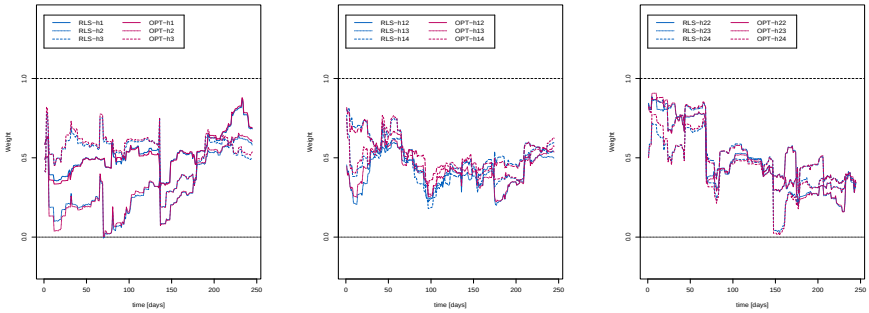


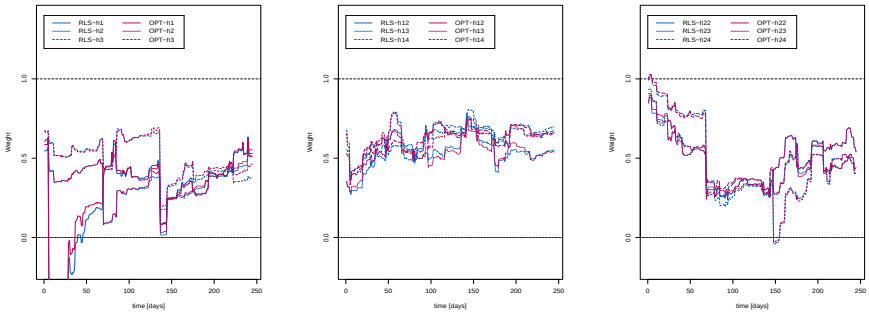
Figure 5.10: Time-varying intercepts in RLS estimation.

bination. All panels show that initially the intercepts are distributed over the origin, but in the end almost all intercepts are below zero or close to zero. There are similarities to the constants in the off-line estimation, illustrated in Figure 5.5, that is the negative values on the constants. It can not be concluded that the intercept is a constant in time when aggregating forecasts, the intercepts are variating throughout the whole data set. It is only in the last days that some stability is reached but the quantity of the intercepts is very high.

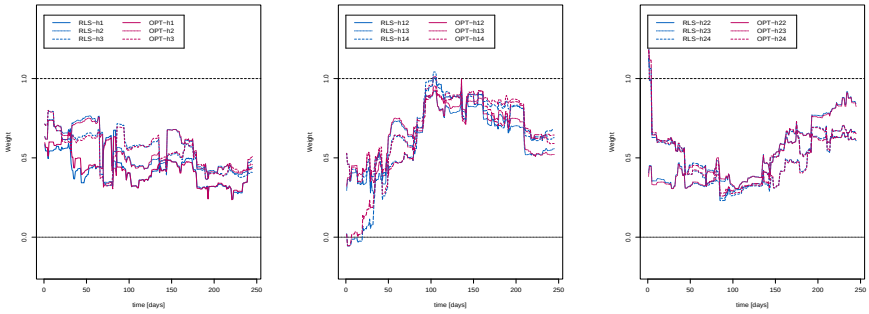
In Figure 5.11 the time-varying weights for RLS method and OPT method are



(c) DWD weights when combined with HIRLAM



(c) DWD weights when combined with MM5



(c) HIRLAM weights when combined with MM5

Figure 5.11: Comparison between RLS and OPT method for time-varying weights.

compared for combining of two predictions. Each row in this 3×3 figure matrix is for each combination and the columns are three levels of prediction horizons: short, intermediate and large with horizons 1-3, 12-14 and 22-24 respectively. The figures show that there is a difference in these two methods in all combinations for all horizon levels. The theory says that without the constant term in the model these methods sould give the same results. Indication of constant is needed as Figures 5.10 and 5.11 illustrate.

5.5.3 Performance of adaptive estimation

With three individual forecasts there are 3 optional combinations of compositing two forecasts. Figure 5.12 shows the results from combining two forecast

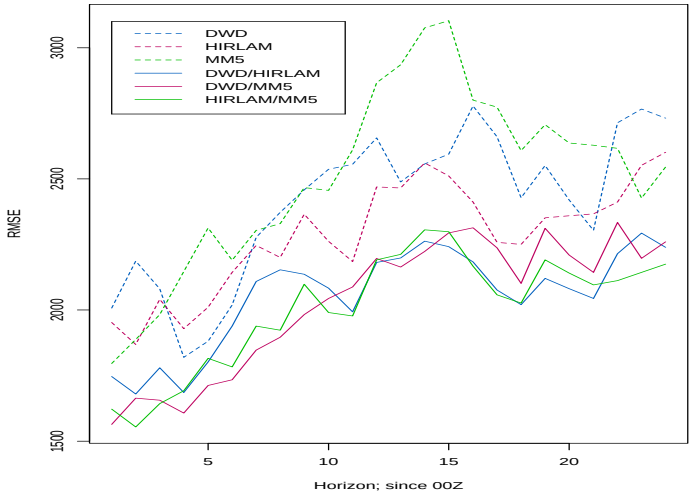


Figure 5.12: RMSE for combination of two forecasts with RLS method, compared to performance of the individual forecasts.

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 in the combination. If the intercept appears to be close to zero it could be neglected and the optimal method would perform as well as the adaptive regression. What Figure 5.13 illustrate 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

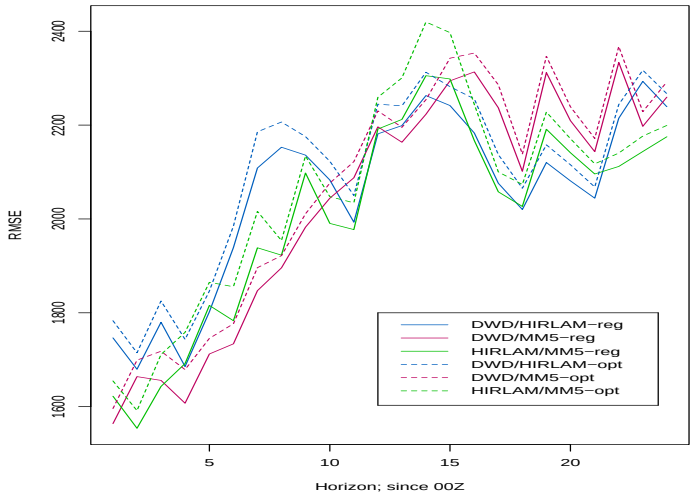


Figure 5.13: Performance comparison between RLS and OPT method when two forecasts are combined.

not be ignored in any of the three possible synthesis.

In Figure 5.14 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 gain is the greatest 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 least squares method when the third prediction is augmented to the composition.

By comparing Figures 5.6 and 5.14 the importance of estimating the weights adaptively is visualized. Great improvement in accuracy is achieved along with the ability of detecting abnormal behaviour in the time-varying weights.

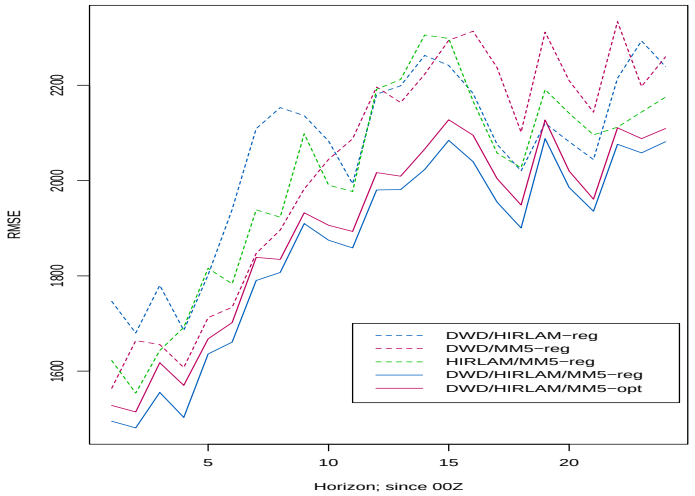


Figure 5.14: Two methods of combining three wind power forecasts compared with RLS method for two prediction combined.

Table 5.2 shows a coefficient of determination for selected horizons for three different methods. It illustrated the superiority 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 5.1 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. To be able to improve accuracy a forecast which appear on the opposite side of the observed production is needed to approach the observations. Forecast errors on either side of the power production would reduce the correlation. The correlation between the forecast errors in Figure 5.2 show the DWD/MM5 having the smallest correlation over the intermediate horizons. The combination of these two concluded in the best combined forecasts with two constituent forecasts.

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

Table 5.2: 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.

Chapter 6

Fitting weights with local regression

6.1 Introduction

Locally weighted regression (section 3.2) is a procedure for fitting a regression surface to data through smoothing. A dependent variable is smoothed as a function of the independent variables in a moving fashion similar to the moving average defined in Section 2.8. For each fitting point on the surface some fraction of the data set is used to estimate the fit where the fraction have 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.

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.

In this chapter the local regression model is shortly introduced in section 6.2. To be able to estimate the weights by local regression the bandwidth have to be properly estimated. This is illustrated in Section 6.3. Finally the local fit is

compared with the RLS method from previous chapter where few bandwidths are inspected.

6.2 Locally fitted weights

When the weights are fitted locally the fitting points are the time steps from the recursive estimation in the previous chapter. For each fitting point in the data the local regression model is

$$y_{c,\tau} - y_{3,\tau} = w_0(\tau) + w_1(\tau)\,(y_{1,\tau} - y_{3,\tau}) + w_2(\tau)\,(y_{1,\tau} - y_{3,\tau}) \tag{6.1}$$

where τ is an index for the fitting point. By substituting $y_{\cdot,t} - y_{3,t}$ for $\tilde{y}_{\cdot,t}$ the model is written as

$$\tilde{y}_{c,\tau} = w_0(\tau) + w_1(\tau)\tilde{y}_{1,\tau} + w_2(\tau)\tilde{y}_{2,\tau}. \tag{6.2}$$

or with a matrix notation it is

$$\tilde{y}_\tau = \tilde{\mathbf{y}}_\tau^T \mathbf{w}(\tau). \tag{6.3}$$

The local model is the local constant model and $\mathbf{p}_0(\tau) = 1$ implies that $\mathbf{z}_\tau \tilde{\mathbf{y}}_\tau$ and the linear model which is fitted locally in (3.7) is simply

$$\tilde{y}_\tau = \boldsymbol{\phi}_{\tau,1}\tilde{\mathbf{y}}_{1,\tau} + \mathbf{e}_\tau \tag{6.4}$$

where weighted least squares is use to estimate the parameters in the local model.

6.3 Selecting the bandwidth for the local fit

In the RLS estimation in chapter 5 the forgetting factor was chosen to represent the past days for the present estimation. The bandwidth is very similar factor, its quantity implies how many data points are used in estimation. Fitting the weights locally is not an adaptive procedure and uses $\alpha/2$, with α denoting the bandwidth, in each direction in estimation. Referring to the bandwidth in the following paragraphs implies the half of the bandwidth.

The selection of bandwidth in smoothing 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 implies that when the bandwidth is small, the residuals are close to zero and the MSE as well. Expanding the bandwidth increases the error terms until it approaches the mean observed value. This is illustrated in Figure 6.1 for the DWD/MM5 forecast. Each prediction horizon has 303 observations which

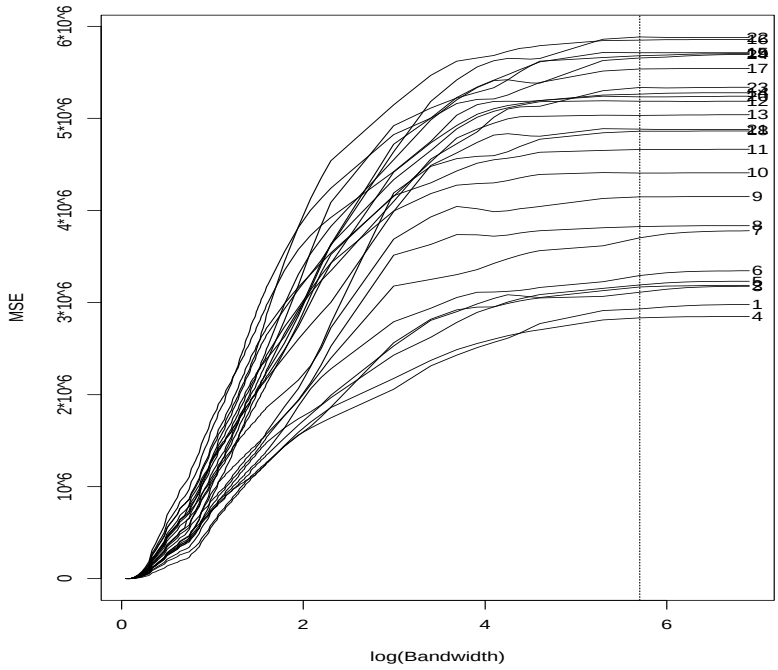
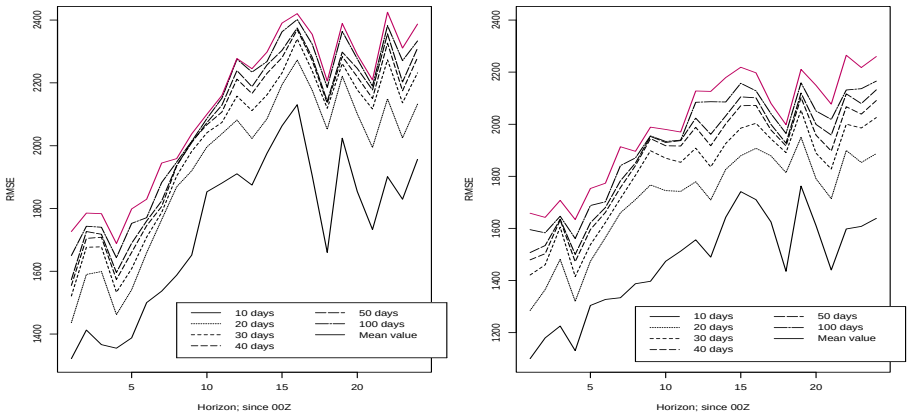


Figure 6.1: Mean squared error as a function of the bandwidth. For bandwidth zero the local estimate is the observed values. With more points used to estimate some fit, departure from the observations is apperant.

implies that the mean value is reached when the bandwidth is 303 days. The dotted horizontal line shows that level in the figure.

The same increase is illustrated in Figure 6.2, for each horizon RMSE increases with increacing bandwidth. The red lines in the panels are the mean values and is the upper limit for the bandwidth. These values are the ones estimated for



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

Figure 6.2: For increasing bandwidth, RMSE for the local fitting 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.

the off-line estimation in section 5.4.

The panels in Figure 6.2 also show how rapidly the RMSE increases for lower bandwidths. Around $\alpha = 40$ (days) the rise almost vanish and the addition of single day to the bandwidth, gives little extension to the performance of the fit.

6.4 comparison with RLS

The RLS method estimates the parameters in the linear model by considering the past data. The future values should also influence an estimation but that is an unknown quantity, which explains the abdrupt changes in the time-varying weights. By estimating the weights with locally weighted regression over the entire data set the data in each direction of the fitting point are used and a smoothed trace for the weights through the data set is detected.

The forgetting factor in RLS estimation was approximated 50 days and by estimating the weights over equal number of days, the bandwidth is initiated at 25 days in each direction from the fitting point. In Figure 6.3 the local fit with

bandwidth of 25 days in either direction, is compared with RLS method. Quite

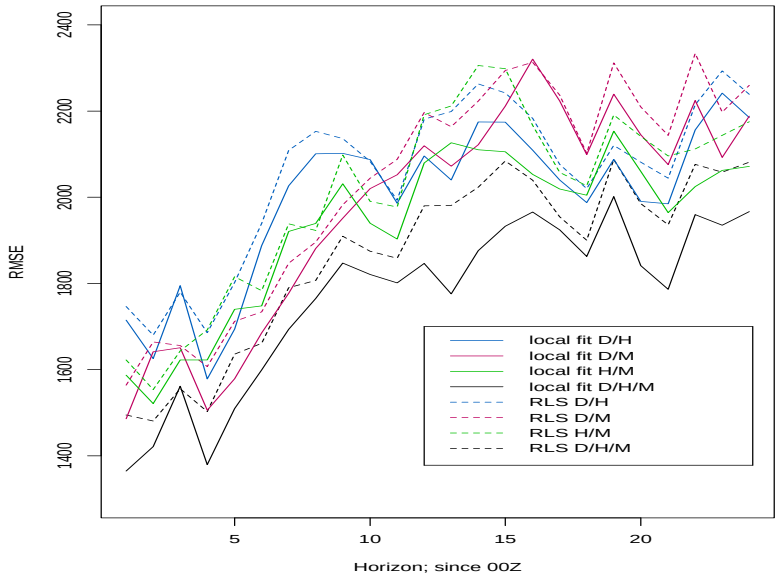


Figure 6.3: RMSE for $\alpha/2 = 25$ compared to RLS.

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, but for the three hours horizon the increase in RMSE from the locally fitted regression exceeds the RLS performance.

Figure 6.4 four horizons from the DWD/MM5 combination are displayed to illustrate how the local fit proceeds compared to the time-varying weights from RLS method. By using data close to a fitting point instead of only considering the previous information, a phase error in the recurively estimated weights is filtered out and more accurate estimation is observed. The phase error can be seen by comparing the weight estimation from RLS and the locally fitted weights.

Appendix C shows this comparison where the bandwidth is selected as $\alpha/2 = 30$. For all weight the phase error is apperant where changes in the parameters are detected upto 25 days ahead. It is also noticed from the Figures in C that with bandwidth 30 in each direction from the fitting point, the sensitivity of the changes is quite high. Small increase in RLS estimations correspond to gross increase in local fit. Expanding the bandwidth reduces the changes in every step of the local fit, but by inflating the bandwidth accuracy of the estimation

is sacrificed.

Figure 6.4 is an example on how the fitted weights adjust with increase in bandwidth. The increase in the bandwidth reduces the amplitude of the local

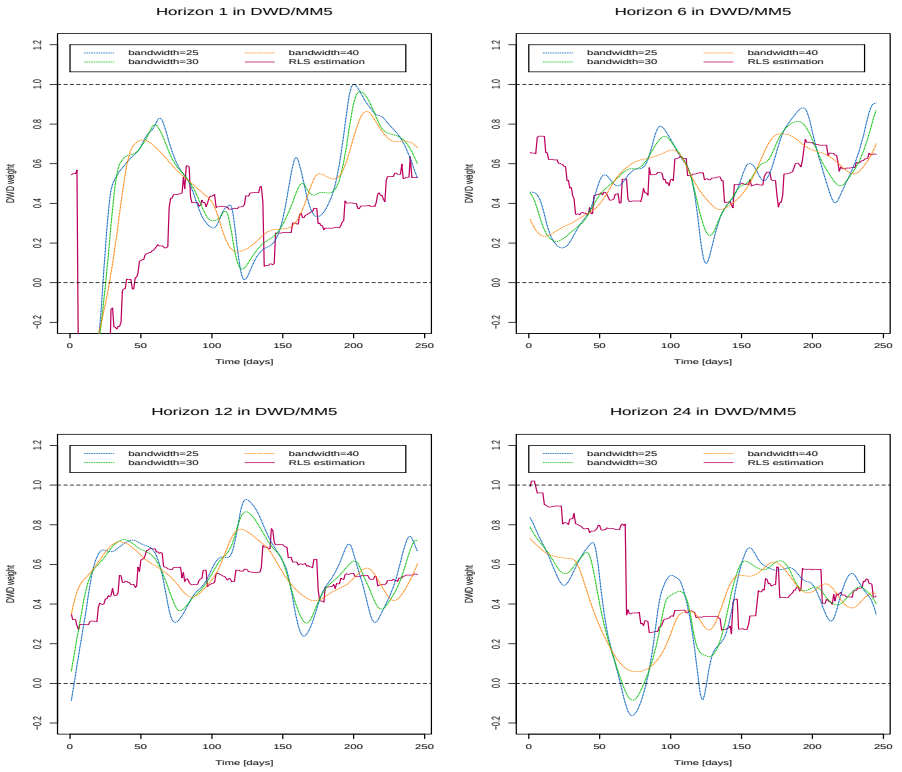


Figure 6.4: Few examples to illustrate how the bandwidth changes compared with the weights from RLS.

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 40×2 appears to be quite alike where the phase error apparts the estimations. For this estimation the local fit spans about 1/4 of the data set for each fitting point.

By using local regression with bandwidth of 40 on either side of the fitting point, the recursive estimation is not outperformed as Figure 6.5 indicates. The performance appears to be similar The performance appears to be similar for the

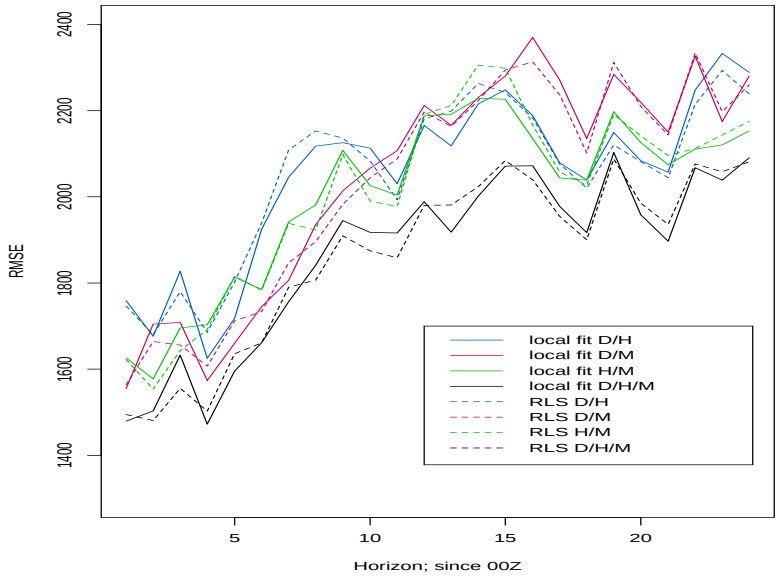


Figure 6.5: RMSE for $\alpha/2 = 40$ compared to RLS.

methods within all combinations. The local regression is wanted to give better performance than RLS estimation which indicates that the bandwidth has to be reduced. For $\alpha/2 = 30$ the local fit is improved such that it outperforms the recursive method in almost all prediction horizons, only the three hours horizon is unique. This is depicted in Figure 6.6.

Table 6.1 shows the coefficient of determination for the local fit with bandwidth 60 days and 80 days, compared with the RLS fit from Table 5.2. 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. The only exception is prediction horizon 3, where the bandwidth need to be reduced further or down to 50 days.

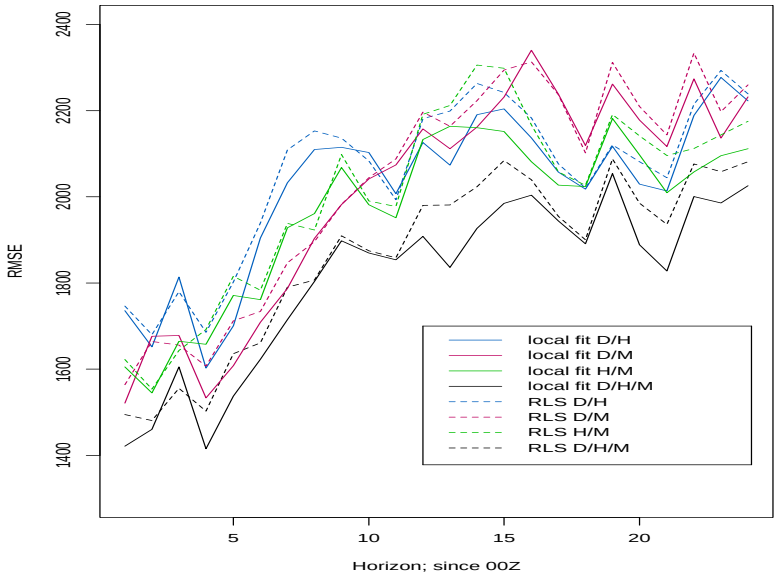


Figure 6.6: RMSE for $\alpha/2 = 30$ compared to RLS.

Combination		Prediction horizon [hours]						
		1	2	3	6	12	18	24
$\alpha/2 = 30$	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
$\alpha/2 = 40$	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

Table 6.1: Comparison between two locally fitted procedures and RLS performances.

Chapter 7

Weight estimation using MET forecasts

7.1 Introduction

The objective in this presentation is to estimate weights in combined forecasts with informations from the MET forecasts. In previous chapters these weights have been estimated by various methods including local regression. In this chapter the local regrssion 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 (2.2), which will give a conditional parametric model of the combined forecast.

Through the analysis there the focus is on combining two forecasts with the 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.

7.2 Dependency between weights and MET forecasts

7.2.1 Linear model

Getting approriate weights is strongly depending on the bandwidth selection for the local fit. It is also noticed that this dependency is related to the prediction horizon. To find weights which fits the MET forecasts, the relation between the bandwidth α and some performance measure, in this case the coefficient of determination, is investigated.

This investigation is performed by fitting a linear model of the MET forecasts to some estimated weights from the local regression with various bandwidth. The intention is to evaluate the bandwidth which gives the best fit. The linear model is the general linear model explained in equation (2.16) with different notation, or

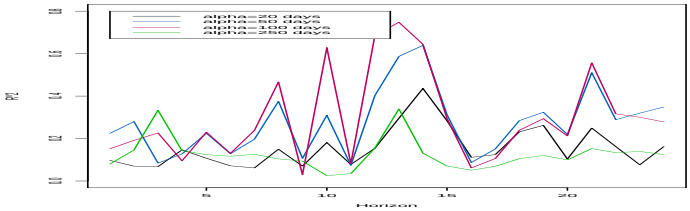
$$\mathbf{w} = \boldsymbol{\beta}^T \mathbf{X} + \mathbf{e} \tag{7.1}$$

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 $\sigma_e^2 = 1$.

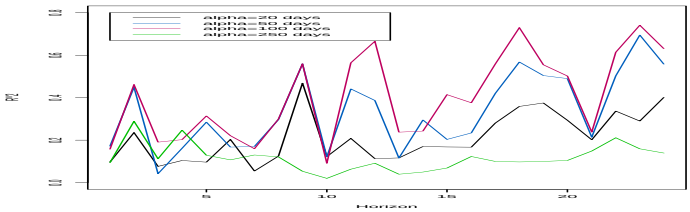
Figure 7.1 shows how the performance changes with different value in the bandwidth, α. The reason for only one weight from each combination is displayed is because the weights are linearly related by the restriction $\sum_{i=1}^{2} w_i = 1$. Four different bandwidths are plotted in each panel and show that with increase in bandwidth to some extend, the performance is improved for the linear model in (7.1). Of the four bandwidths $\alpha = 100$ gives the best fit for the model. With bandwidth larger than 100 days the performance reduces until it reaches some limit when the bandwidth spans the entire data set. Even with 100 days presenting the maximum R^2, bandwidth of 50 days is selected for further analysis with the MET forecasts. From the panels in Figure 7.1 it is observed that the gain of double the bandwidth is not significant and therefore is $\alpha = 50$ chosen for the locally fitted weights to generate the dependency to the MET forecasts.

7.2.2 Partitioning MET variables

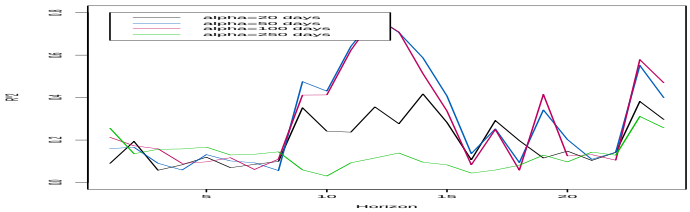
Through the analysis with the MET variales the main focus is on using the DWD/HIRLAM combination as basic synthesis. Other two combinations are



(a) DWD/HIRLAM



(b) DWD/MM5



(c) HIRLAM/MM5

Figure 7.1: Coefficient of determination for various bandwidth in all horizons. Increase in bandwidth gives increase in R^2 to 100.

analyzed as well but not graphically displayed.

By Inspecting the scatterplots in Figure 7.2 it is difficult to see trends between the DWD weight and the MET forecasts. The red line in the plots is locally weighted regression between corresponding MET variable and the DWD weights. The weights seems 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 it only shows coherency with one variable. Relation of response variable with two explanatory variables can be demonstrated by *coplots* which is well illustrated in [3] in relation with conditionally

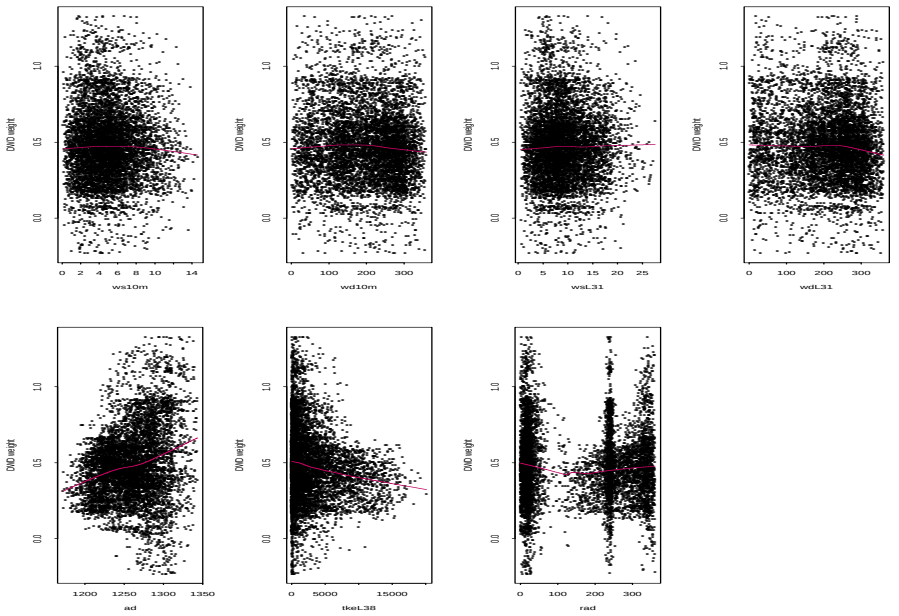


Figure 7.2: Scatterplot of the DWD weight in DWD/HIRLAM in relation with the MET forecasts.

parametric fits. One explanatory variable is partitioned in 2-4 categories and the response variable is smoothed w.r.t. the other explanatory variable within the partitioning. The coplots for the DWD weights are diplayed in appendix D, but Figure 7.3 show some of the coplots where air density (**ad**) and/or turbulent kinetic energy at level 38 (**tke**) are one or both of the explanatory variable. 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 parameters in the conditional parametric model

$$\hat{y}_c = w_0(\mathbf{ad}, \mathbf{tke}) + w_1(\mathbf{ad}, \mathbf{tke})\hat{y}_1 + w_2(\mathbf{ad}, \mathbf{tke})\hat{y}_2 \tag{7.2}$$

with the constraint of forecast coefficients summing to one.

The scatterplot of **ad** and **tke** makes a basis for the weight estimation in (7.2) but by observing the **ad**/**tke**-plane in Figure 7.4 it shows the MET data not covering the whole plane. The **tke** forecast is more densed at low values and then diffuses when it increases, while the **ad** forecast is closer to be normally distributed. The distributions are shown in Figure 4.5 in the data description. The data disperse with increase in **tke** and increasing distance from mean value of the air density. This implies that for high values of **tke** and either high or low

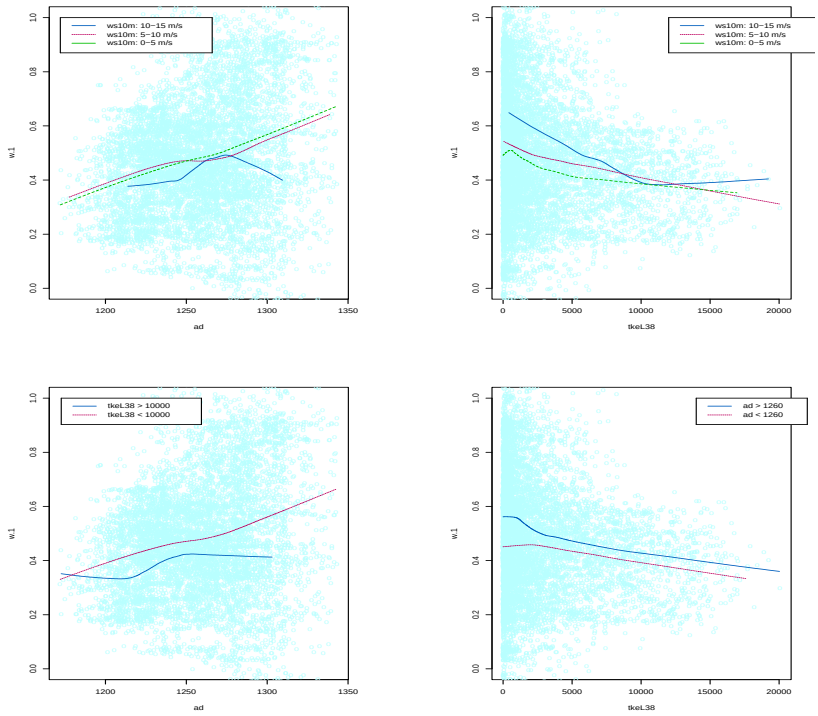


Figure 7.3: Coplots with at least air density or turbulent kinetic energy as one of the explanatory variables

values of **ad**, no or few observed values exist. The convex hull[1] is thus defined as the points inside the red lines in Figure 7.4, that is the area where the weights have some valid estimation on the basic plane.

Figure 7.5 shows the time series for the two MET forecasts which are of interest. The upper panel shows the air density from February to beginning of December. Air density is known to follow the behaviour 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 wether phenomenas of the atmosphere. It varies through the entire period, less though in both tails which indidates reduced variation over the winter period. **tke** is a variable used to study turbu-

[1]Convex hull or for a set of points X in a real vector space V is the minimal convex set containing X. Information from *wikipedia.org*

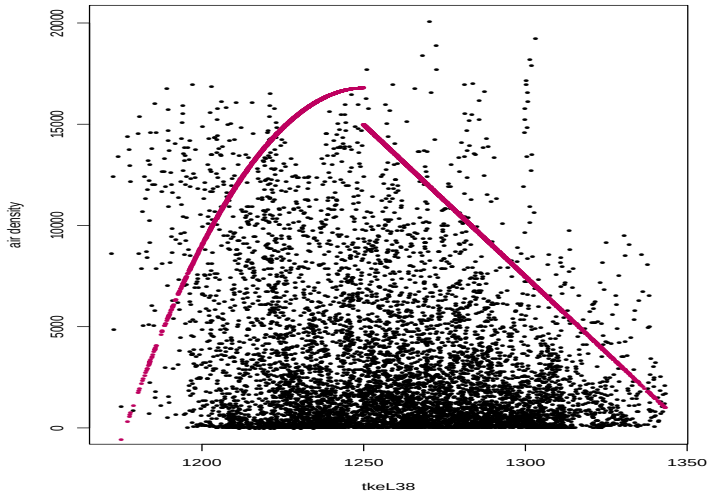


Figure 7.4: Scatterplot of **ad** and **tke** variables. These variables make the basis for the surface of the weight estimation. The red lines define the convex hull.

lence 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.

7.3 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 has been illustrated in section 5.3 with the restriction on the parameters, the model used to estimate the weights can be rewritten as

$$\hat{y}_c - \hat{y}_2 = w_0(\mathbf{ad}, \mathbf{tke}) + w_1(\mathbf{ad}, \mathbf{tke})\,(\hat{y}_1 - \hat{y}_2) \tag{7.3}$$

when combining two individual forecasts and denoted 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

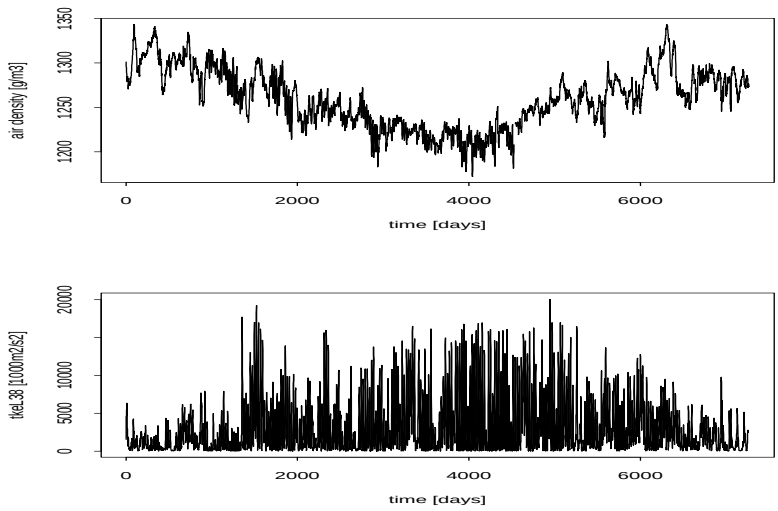


Figure 7.5: Time series plots for the air density and the turbulent kinetic energy at level 38

w_0 a local constant is approximated. The local models are

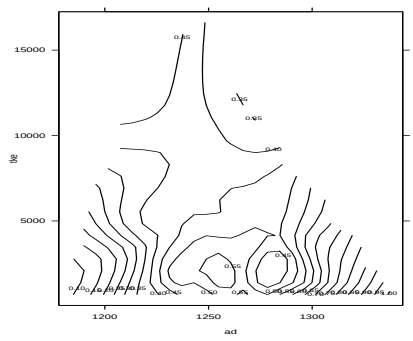
$$w_0(\mathbf{ad}, \mathbf{tke}) = w_0 \tag{7.4}$$

$$w_1(\mathbf{ad}, \mathbf{tke}) = w_{10} + w_{11} \cdot \mathbf{ad} + w_{12} \cdot \mathbf{tke}. \tag{7.5}$$

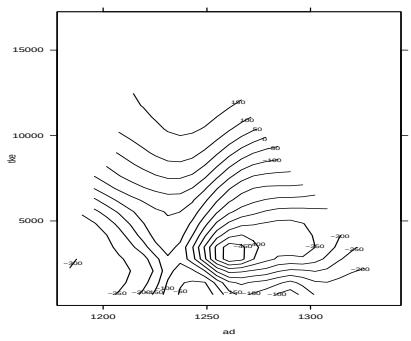
By substituting (7.4) and (7.5) into (7.3) a modified linear model for combination is denoted:

$$\begin{aligned} \hat{y}_c - \hat{y}_2 = w_0 &+ w_{10}\,(\hat{y}_1 - \hat{y}_2) \\ &+ w_{11} \cdot \mathbf{ad}\,(\hat{y}_1 - \hat{y}_2) \\ &+ w_{12} \cdot \mathbf{tke}\,(\hat{y}_1 - \hat{y}_2)\,. \end{aligned} \tag{7.6}$$

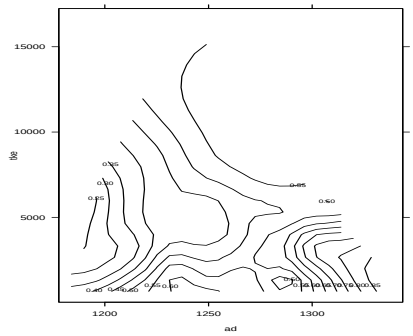
where the explanatory variables are now products of prior variables and the MET forecasts. The modified model in 7.6 now includes new explanatory variables, which are the products, and instead of estimating two parameters four are evaluated in the modified formulation. Combining wind power forecasts as in (7.6) 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 7.6 are obtained. The local fit in the analysis uses the *nearest neighbor* spanning 2/3 of the data set.



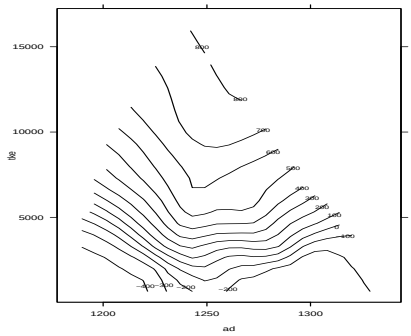
(a) Weight for DWD in D/H



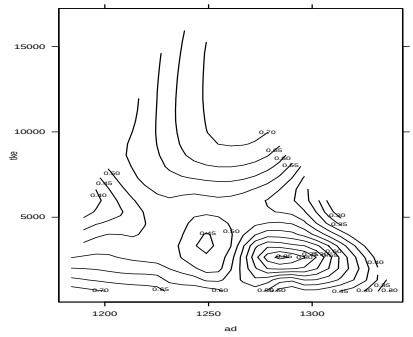
(b) Intercept in D/H



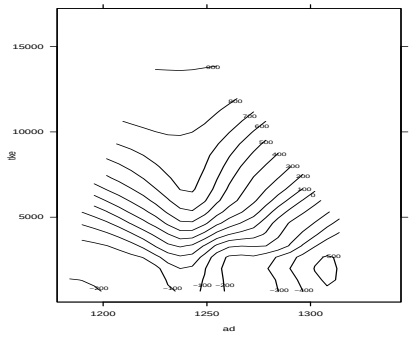
(c) Weight for DWD in D/M



(d) Intercept in D/M



(e) Weight for HIRLAM in H/M



(f) Intercept in H/M

Figure 7.6: Contour plots for weights and the intercepts

The weights on DWD when combined with HIRLAM and MM5 appear to have similar surface, 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 reducing with progressing air density. The behaviour 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.

7.3.1 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:

$$w_1(\mathbf{ad}, \mathbf{tke}) = v_{10}(\mathbf{tke}) + v_{11}(\mathbf{tke}) \cdot \mathbf{ad}, \tag{7.7}$$

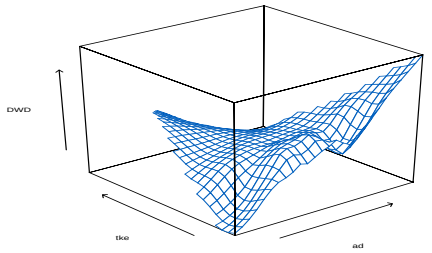
where v_{10} and v_{11} are the intercept and the slope respectively. This can be detected from the contourplot in Figure 7.6(a) or by the surface plot in Figure 7.7 where the approximated linearity can be visualized. The contourplot in shows that for increase in **tke**, the intercept increases but the slope decreases and become negative for **tke**'s higher than 8000.

Observing the intercept $w_0(\mathbf{ad}, \mathbf{tke})$ shows a wavelike behaviour 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 (7.3), e.g. the intercept and the product of the forecast weight and the forecast. Table 7.1 shows the covariance matrix of these terms and

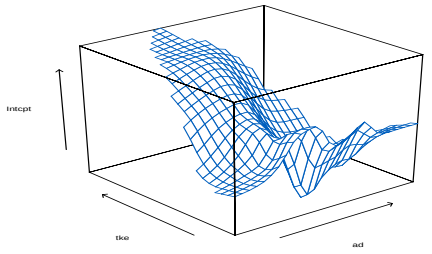
Variance	$w_0(\mathbf{ad}, \mathbf{tke})$	$w_1(\mathbf{ad}, \mathbf{tke})\tilde{y}_1$
$w_0(\mathbf{ad}, \mathbf{tke})$	17369.5	24039
$w_1(\mathbf{ad}, \mathbf{tke})\tilde{y}_1$	24039	1129552

Table 7.1: Covariance matrix for terms in (7.3).

it indicates that the product between the weight and the predictor is about 8



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



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

Figure 7.7: Surface plots for the DWD weight and the intercept

times greater than the variance of the intercept. The variations for low **tke** on the surface of the intercept are therefore omitted.

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**:

$$w_0(\mathbf{ad}, \mathbf{tke}) = v_0(\mathbf{tke}). \tag{7.8}$$

This assumption might be a bit crude but will make it alot easier to implement the estimated functions from (7.7) and (7.8) to a conditional parametric model:

$$\begin{aligned}\hat{y}_c - \hat{y}_2 &= v_0(\mathbf{tke}) + [v_{10}(\mathbf{tke}) + v_{11}(\mathbf{tke}) \cdot \mathbf{ad}]\,(\hat{y}_1 - \hat{y}_2) \\ &= v_0(\mathbf{tke}) + v_{10}(\mathbf{tke})(\hat{y}_1 - \hat{y}_2) + v_{11}(\mathbf{tke})(\hat{y}_1 - \hat{y}_2)\mathbf{ad} \\ &= v_0(\mathbf{tke}) + v_{10}(\mathbf{tke})z_1 + v_{11}(\mathbf{tke})z_2 \end{aligned} \tag{7.9}$$

where $z_1 = \hat{y}_1 - \hat{y}_2$ and $z_2 = (\hat{y}_1 - \hat{y}_2)\mathbf{ad}$. The model in (7.9) 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 7.8 shows how the weights in (7.9) change with **tke**. The same valley appears in for the intercept and the same test is performed as before to estimate sufficiency of the variance of the intercept. Table 7.2 shows the covari-

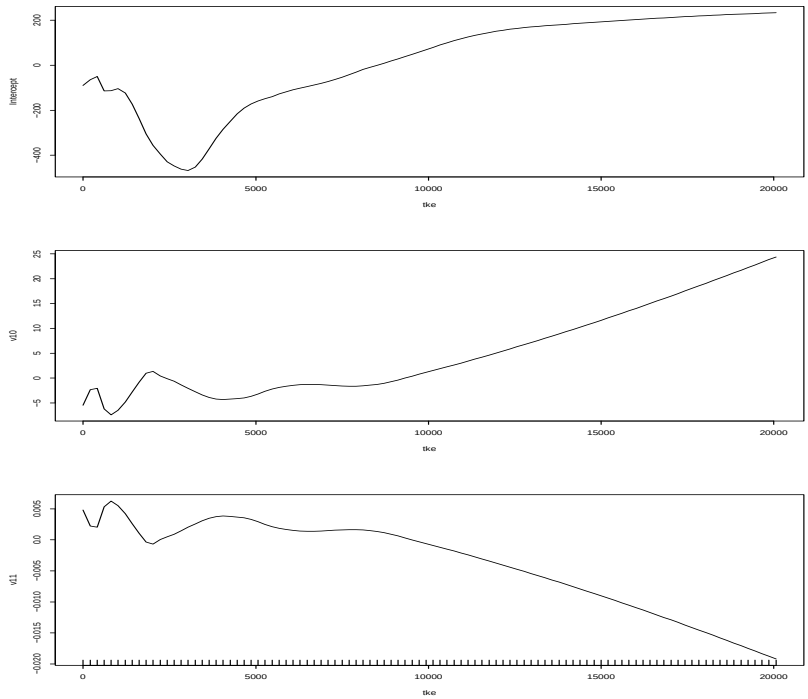


Figure 7.8:

ance matrix of the terms in (7.9) and reveals that the variance of the intercept is only 1/3000 of the other variances and therefore can the valley at **tke** around 3000 be neglected.

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

Table 7.2: Covariance matrix for the terms in (7.9)

The opposite behaviour of the parameters v_{10} and v_{11} is not surprising since they form the weight in the original model 7.3. From Figure 7.8 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.

7.4 Comparison with foregoing methods

The performance for the conditional parametric model is compared with the RLS method and the static model from section5.4. The performances for the surface estimations above are generated by in-sample RMSE and coefficient of determination, which was also performed for the off-line estimation.

In Figure 7.9 the forecasts generated by using MET variables are compared to the off-line 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 has lower RMSE. For larger horizons forecasts using MET variables are outperforming the static model, estimated over the data set.

With MET based forecasts outperforming the off-line performance for large horizons it is interesting to compare it with the RLS performance. This is depicted in Figure 7.10. What this comparison reveals is that over the intermediate prediction horizons the MET dependent forecasts are very close to the RLS performance. For small horizons all the combined forecasts are significantly different. This difference is highest for the smallest 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.

In Table 7.3 R^2 for MET dependent forecasts are compared to the coefficient of determination for the off-line method and RLS method. From the table it

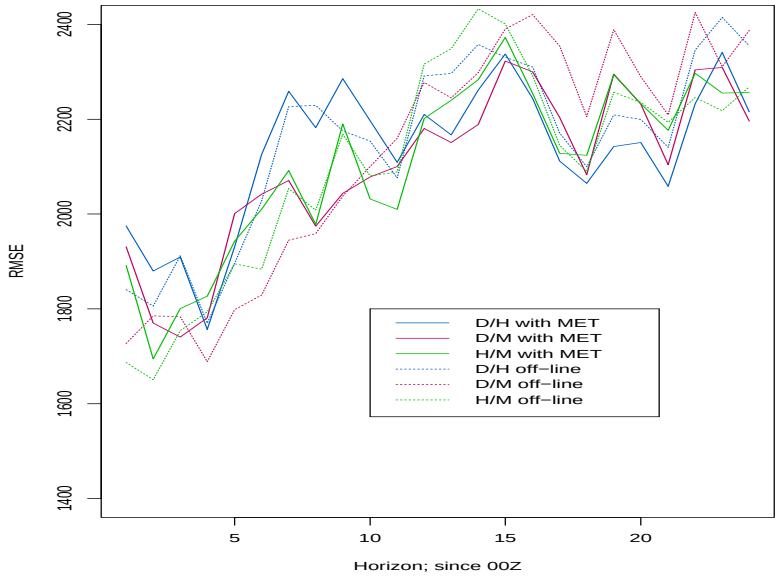


Figure 7.9: Compare MET dependent forecasts with off-line performance.

can be concluded that the fit for MET based forecasts are 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 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.

The performance of the conditional parametric model in (7.9) is depicted in Figure 7.11 with an orange. It appears to be approaching the off-line performance but with a little improvement. Adaptive estimation would even improve it eminently.

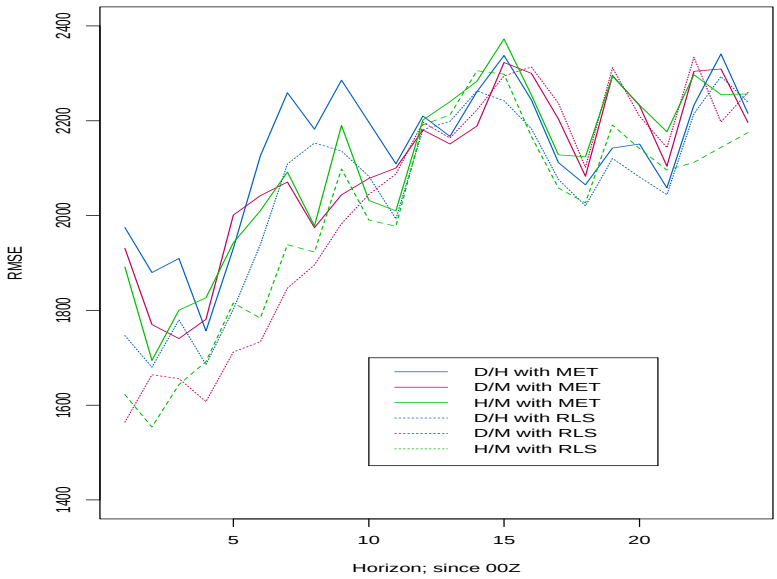


Figure 7.10: Compare MET dependent forecasts with RLS performance.

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
Off-line	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

Table 7.3: Comparing the MET dependent forecasts to the foregoing methods in the presentation.

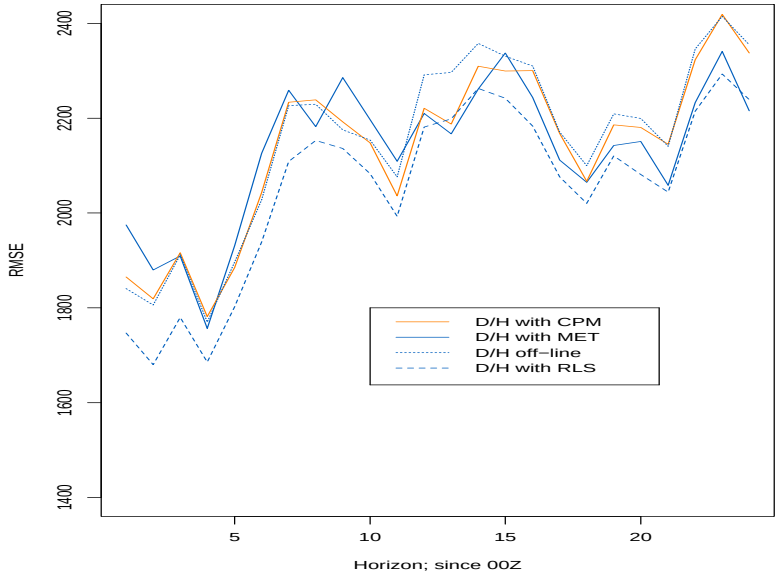


Figure 7.11: Performance of the modified model in (7.9) compared to other DWD/HIRLAM performances.

Chapter 8

Conclusion

The chapter contains a short summary of the results found in the thesis, will be given followed by section about further works related to the thesis.

8.1 Summary of results

Various methods for combining forecasts have been introduced where several were used to combine wind power forecasts adaptively. Of the methods generated, the linear regression model including both constant and restriction outperformed the other methods.

Locally weighted regression improved the recursive least squares method when the bandwidth spaned less than 60 days and the phase error in the weights from the recusive method is detected. The local fit gave the correct estimates for the weights since it concerns all points close to a fitting point, but not only the past data. Using the locally fitted weights to detect informations from the meteorological forecasts, air density and turbulent kinetic energy were extracted. The weights in the linear regression model were conditioning on the two meteorological forecasts. The performance for the weights depending on the MET forecasts gave quite similar results as the linear model estimated for the entire data set in the short predicton horizons. When the horizon was enlarged the

performance improved and was found to be comparable to RLS method.

The surface for the DWD weights appeared to be a linear function of air density where intercept and slope were depending on the turbulent kinetic energy. This was implemented in the linear model which resulted in a sligthly improved performance compared to the off-line procedure.

8.2 Further works

Following this thesis there are few things which can be elevated.

First is the adaptation of the varying coefficient-function developed in chapter 7. The behaviour of the combined forecasts where MET variables are used to generate weights is a bit better than forecasts with parameters estimated for the whole data set. By estimating the weights adaptively extensive improvement is expected.

Combining more than two forecasts by using meteorological variables is also something that cab be considered. In chapter 5 it is concluded that combining all three wind power forecasts give the most accurate combination. The third forecast is not deemed in the thesis but is an interesting topic for further analysis to improve combined forecasts using meteorological variables.

The meteorological forecasts used in the thesis are only the ones from DMI-HIRLAM. Three power forecasts are used in the analysis which all have different meteorological data. Applying more MET forecasts to the analysis gives the researcher a lot more information to work with since the inputs for the weight estimation are the MET forecasts.

The data set used spans only ten months. Since the application in the thesis is related to meteorology where the season is one year, it would be alluring to observe how the MET forecasts affect the individual forecasts in a combination over several seasons.

Appendix A

MET scatterplots and data description

The appendix includes list of the variables available in this study. Some of the variables are highly correlated and are therefore omitted in the analysis. This appendix also includes the scatterplots for the strongly correlated variables.

WPPT forecasts

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

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

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

MET forecasts

ws10m: Forecasted wind speed at 10 meters above ground level (m/s).

wd10m: Forecasted wind direction at 10 meters above ground level (degrees).

rad: Forecasted radiation (W/m^2)

fv: Forecasted friction velocity (m/s).

ad: Forecasted air density (g/m^2).

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

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

tkeL··: Forecasted turbulent kinetic energy in model level ··, the levels in the data set are 31, 38, 39 and 40 (Wm^2/s^2).

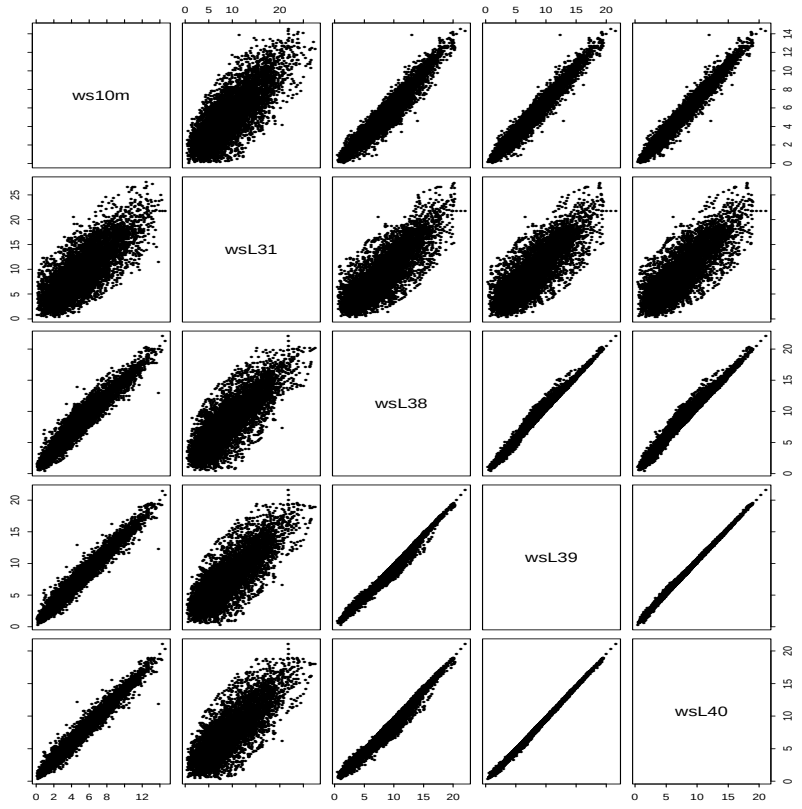


Figure A.1: Pairwise scatterplot of the wind speed variables in the MET data. The correlation is very strong between these variables so only one is used as an explanatory variable in analysis. The wsL31 variable is not as correlated as the other variables and is therefore also included in the data analysis.

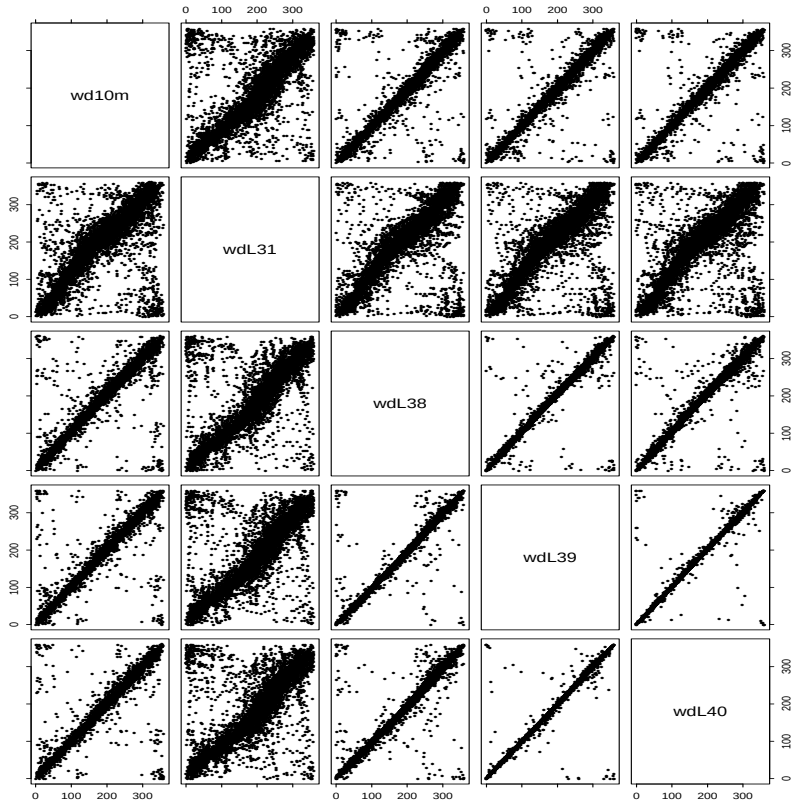


Figure A.2: Pairwise scatterplot of the wind direction variables in the MET data. The correlation is very strong between these variables so only one is used as an explanatory variable in analysis. The wdL31 variable is not as correlated as the other variables and is therefore also included in the data analysis.

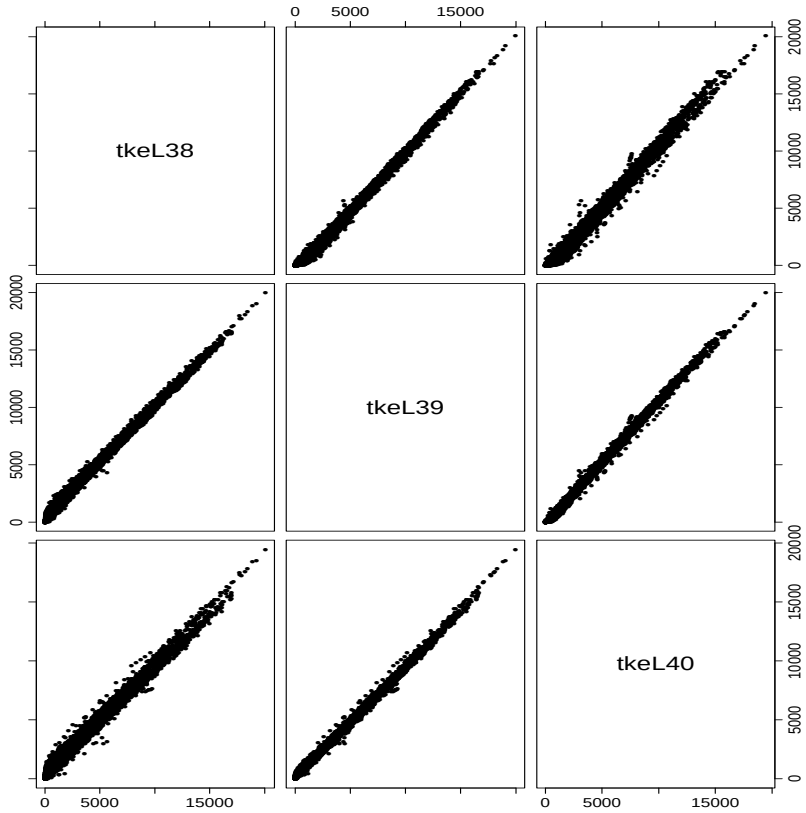


Figure A.3: Pairwise scatterplot of the turbulent kinetic energy variables in the MET data. The correlation is very strong between these variables so only one is used as an explanatory variable in analysis

Appendix B

Time-varying weights from RLS estimation

The appendix includes the the time-varying weights for all combinations from the analysis illustrated in chapter 5.

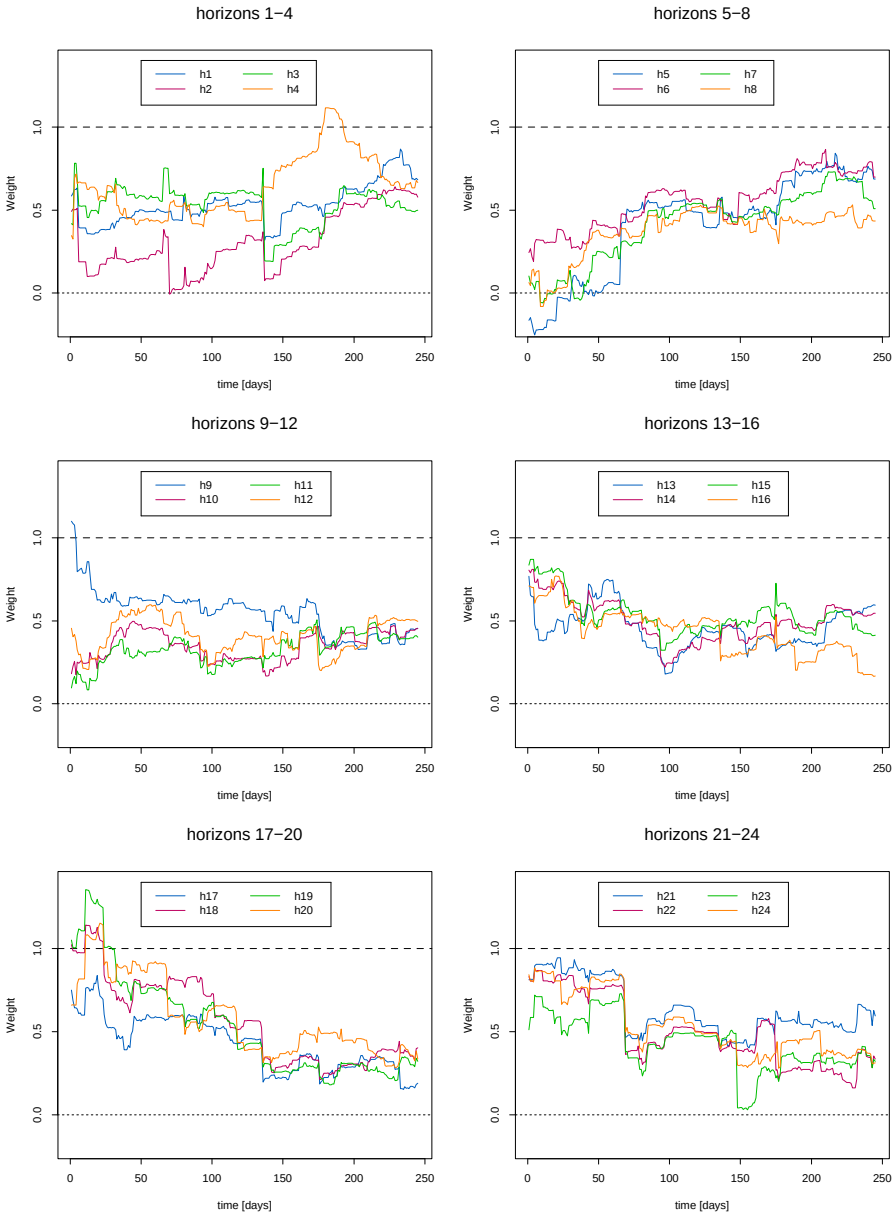


Figure B.1: Weights on DWD forecast when combined with HIRLAM forecast.

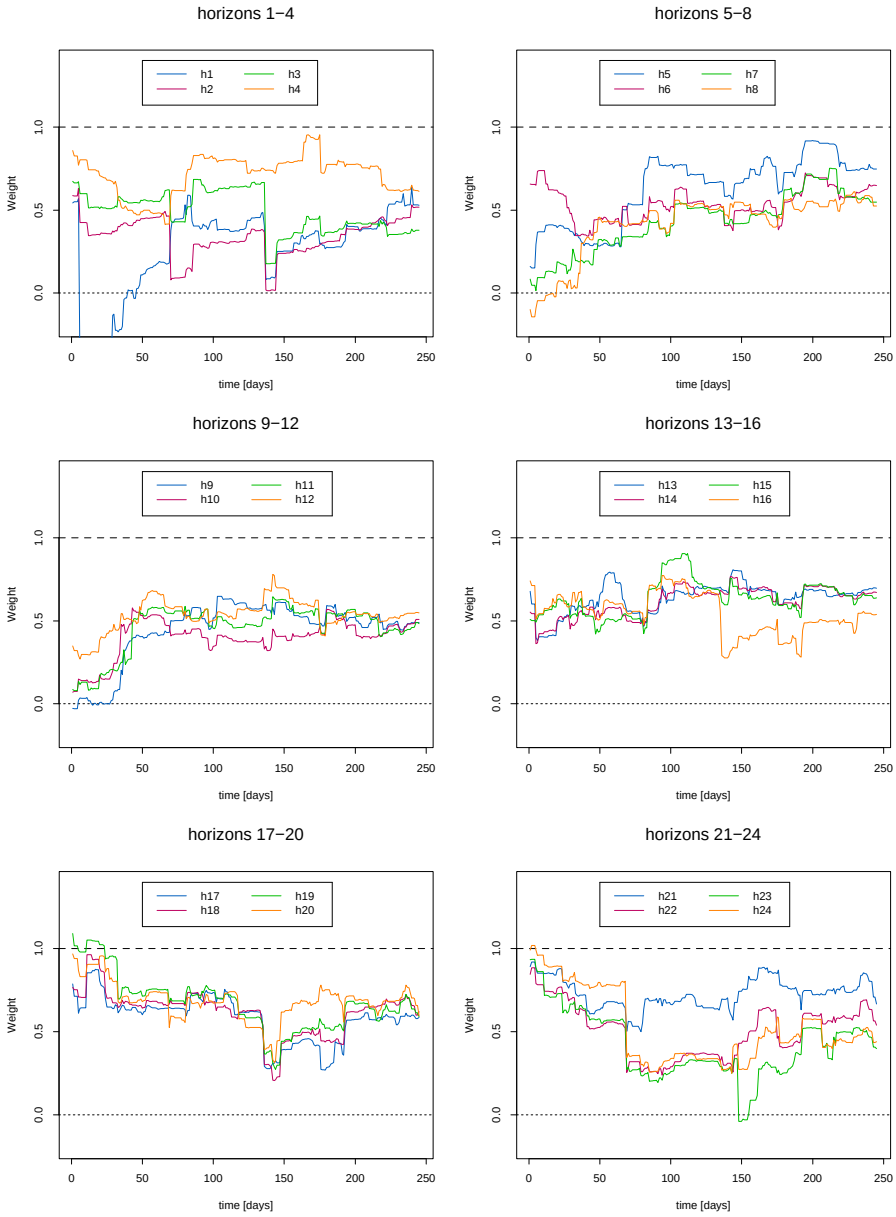


Figure B.2: Weights on DWD forecast when combined with MM5 forecast.

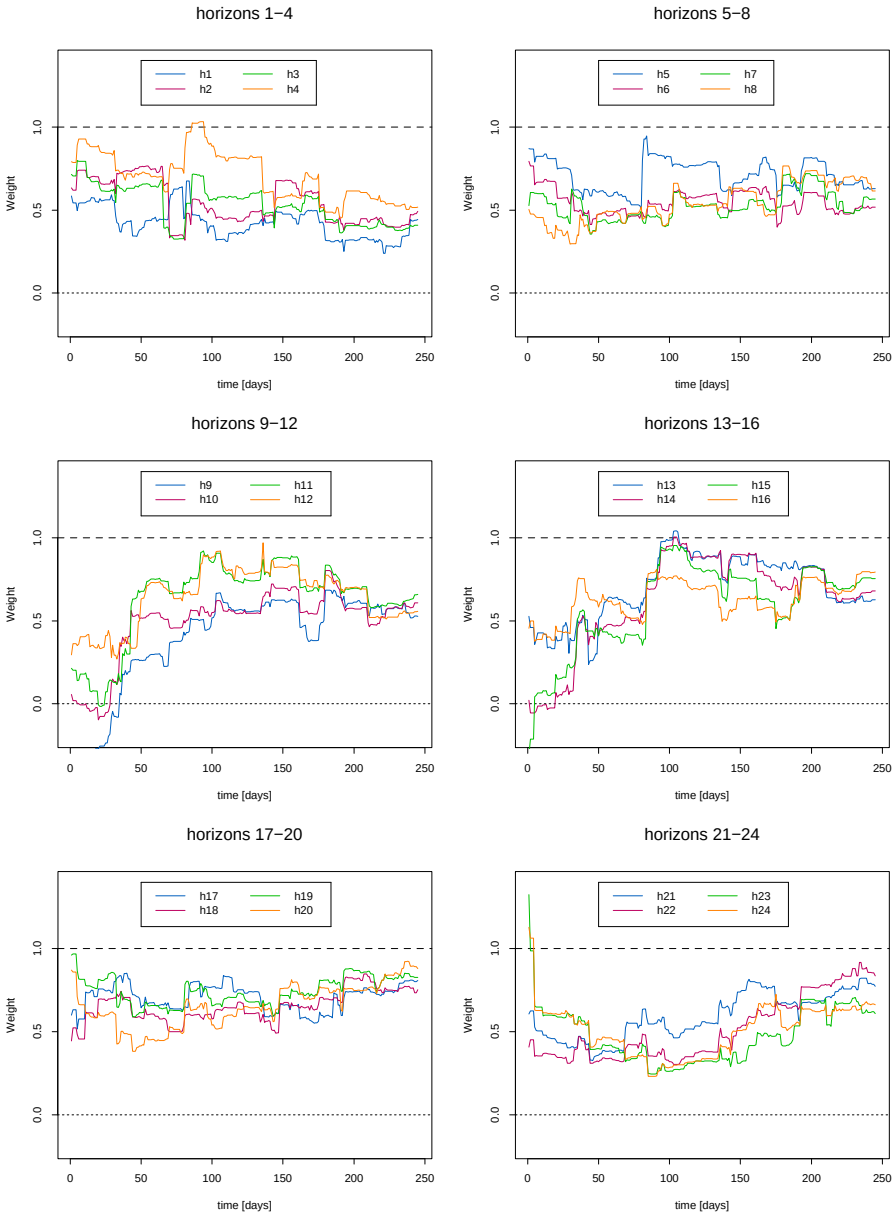


Figure B.3: Weights on HIRLAM forecast when combined with MM5 forecast.

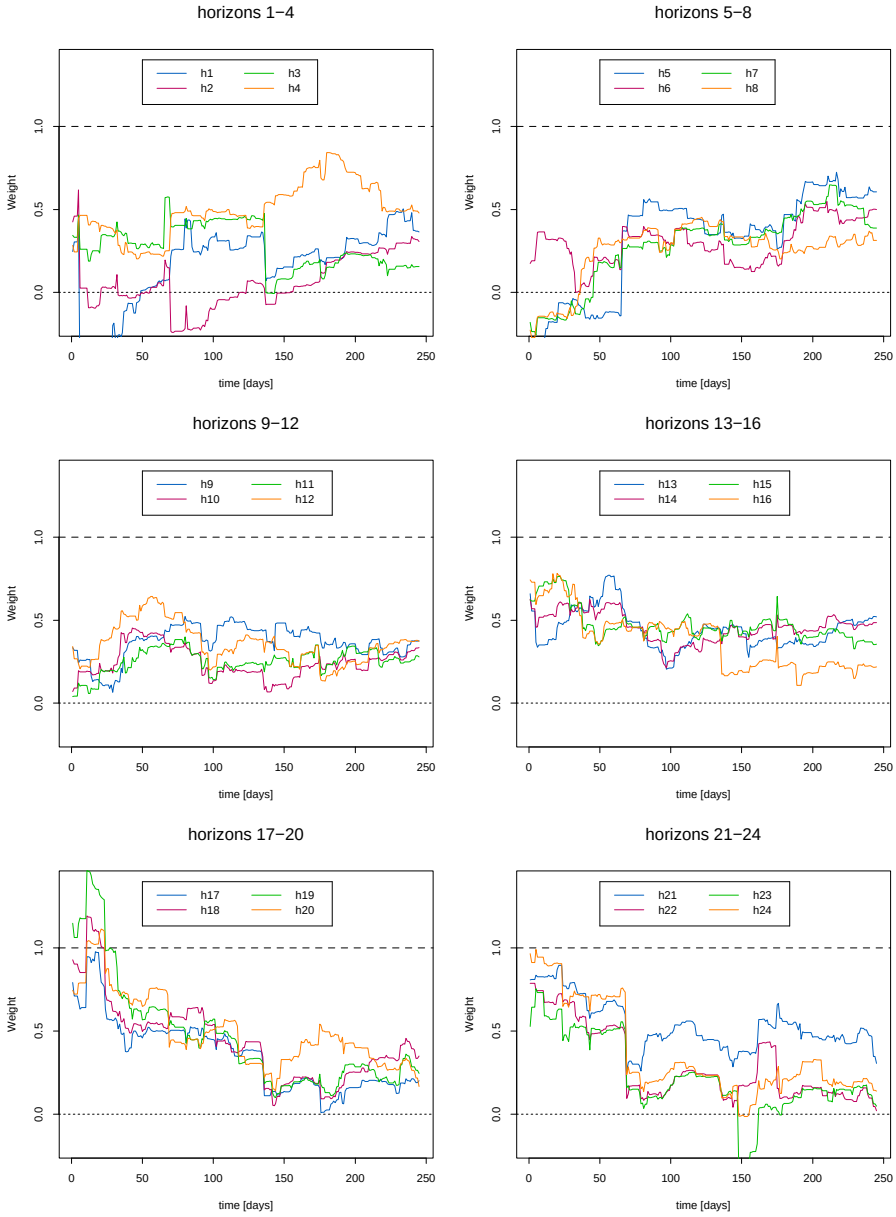


Figure B.4: Weights on DWD forecast in DWD/HIRLAM/MM5 combination.

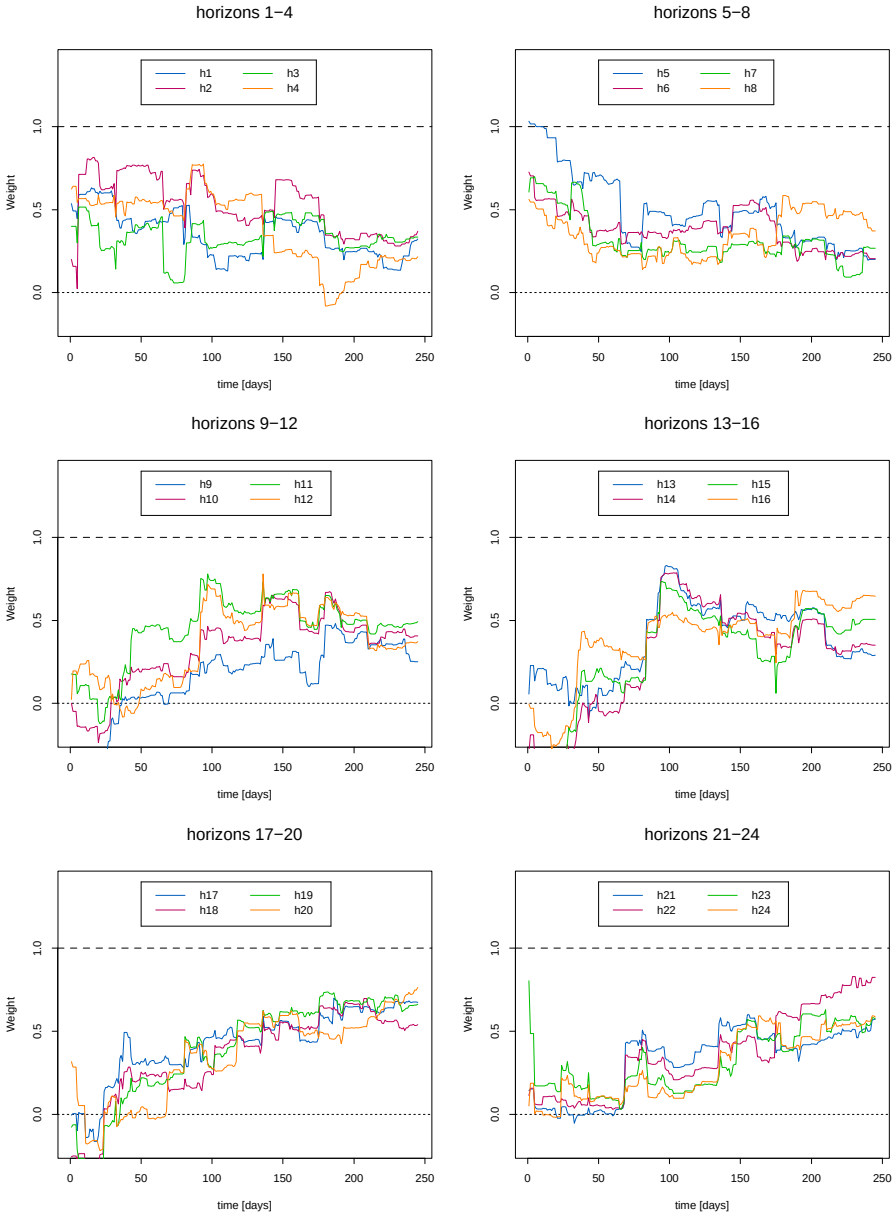


Figure B.5: Weights on HIRLAM forecast in DWD/HIRLAM/MM5 combination.

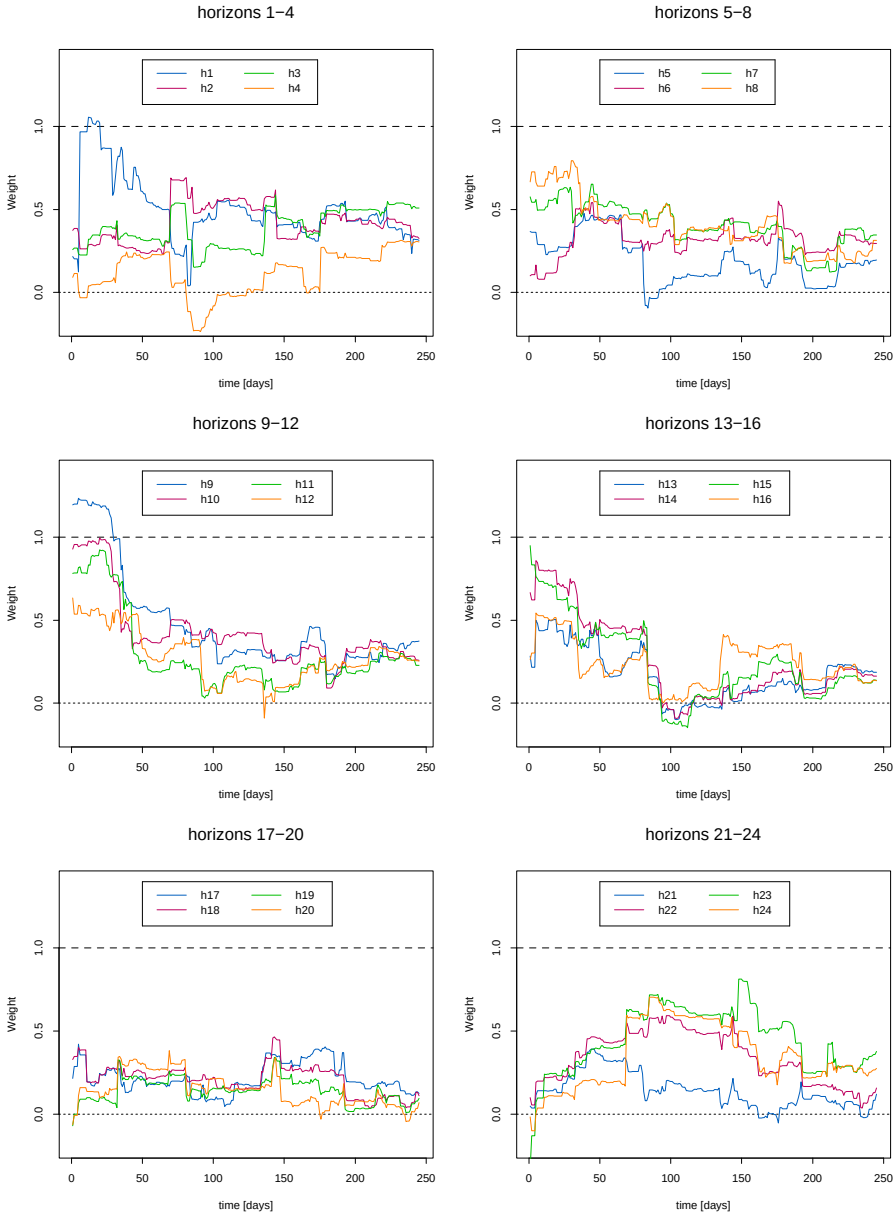


Figure B.6: Weights on MM5 forecast in DWD/HIRLAM/MM5 combination.

Appendix C

plots for locally fitted weights

The appendix includes the local fits for all combinations for bandwidth equal to 60 days. The figures also include the time-varying weights from RLS method for comparison.

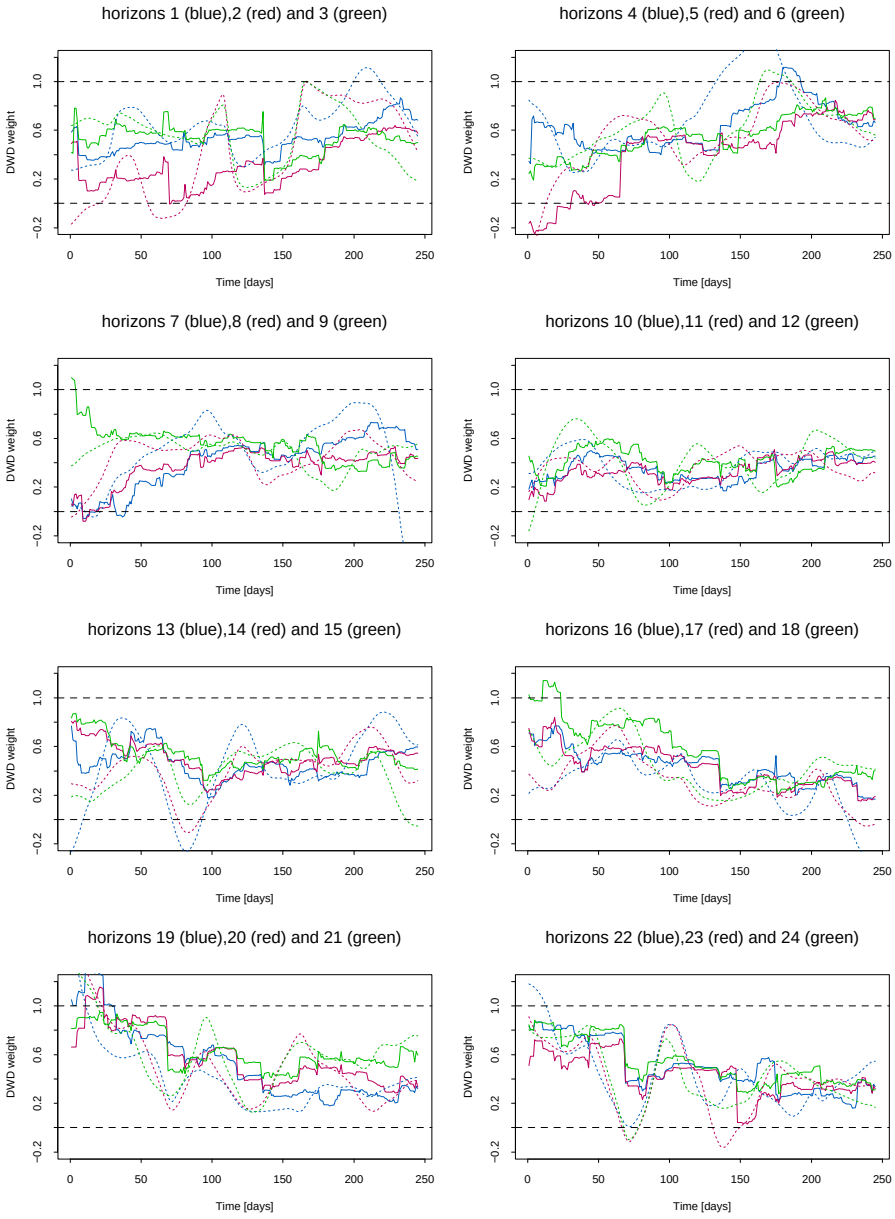


Figure C.1: Local fit (dashed lines) for DWD weights in DWD/HIRLAM, compared with RLS estimations (solid lines) for corresponding combination.

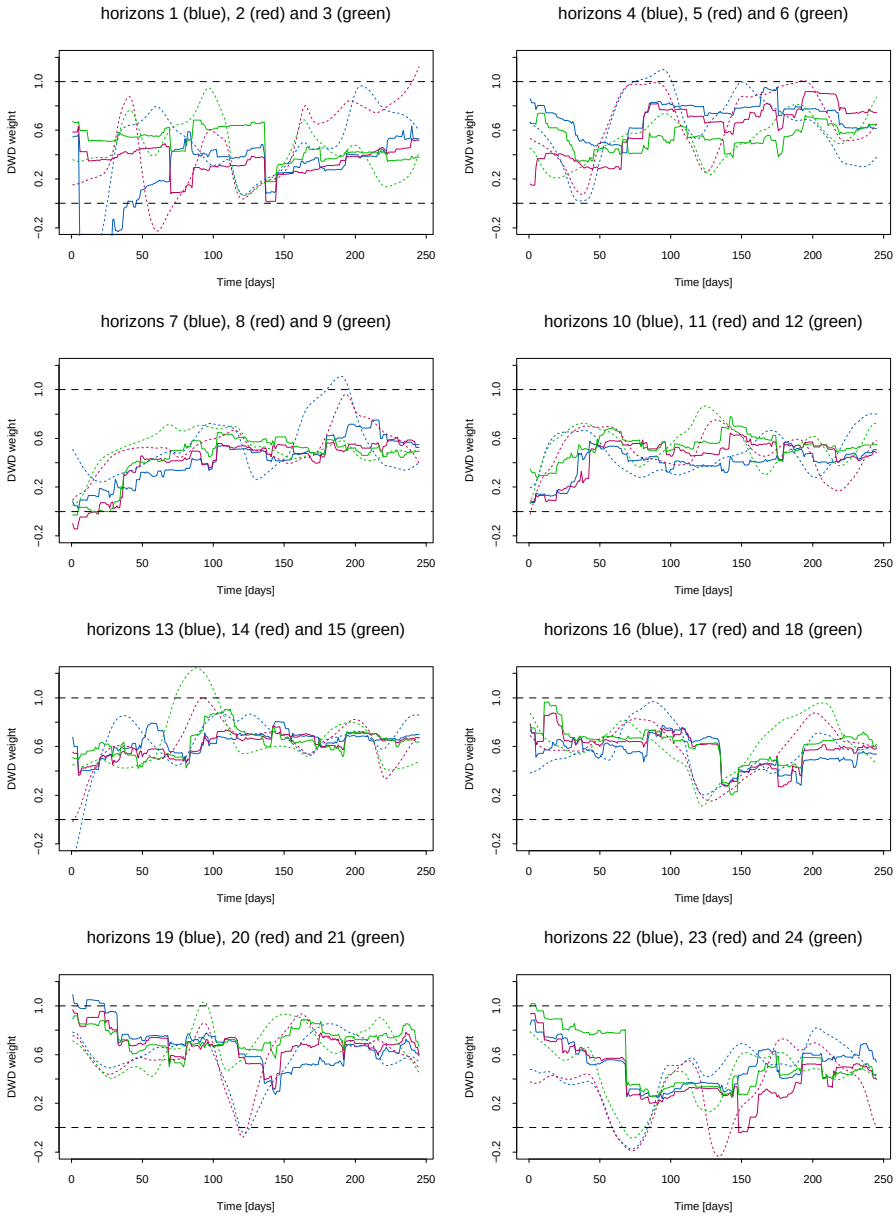


Figure C.2: Local fit (dashed lines) for DWD weights in DWD/MM5, compared with RLS estimations (solid lines) for corresponding combination.

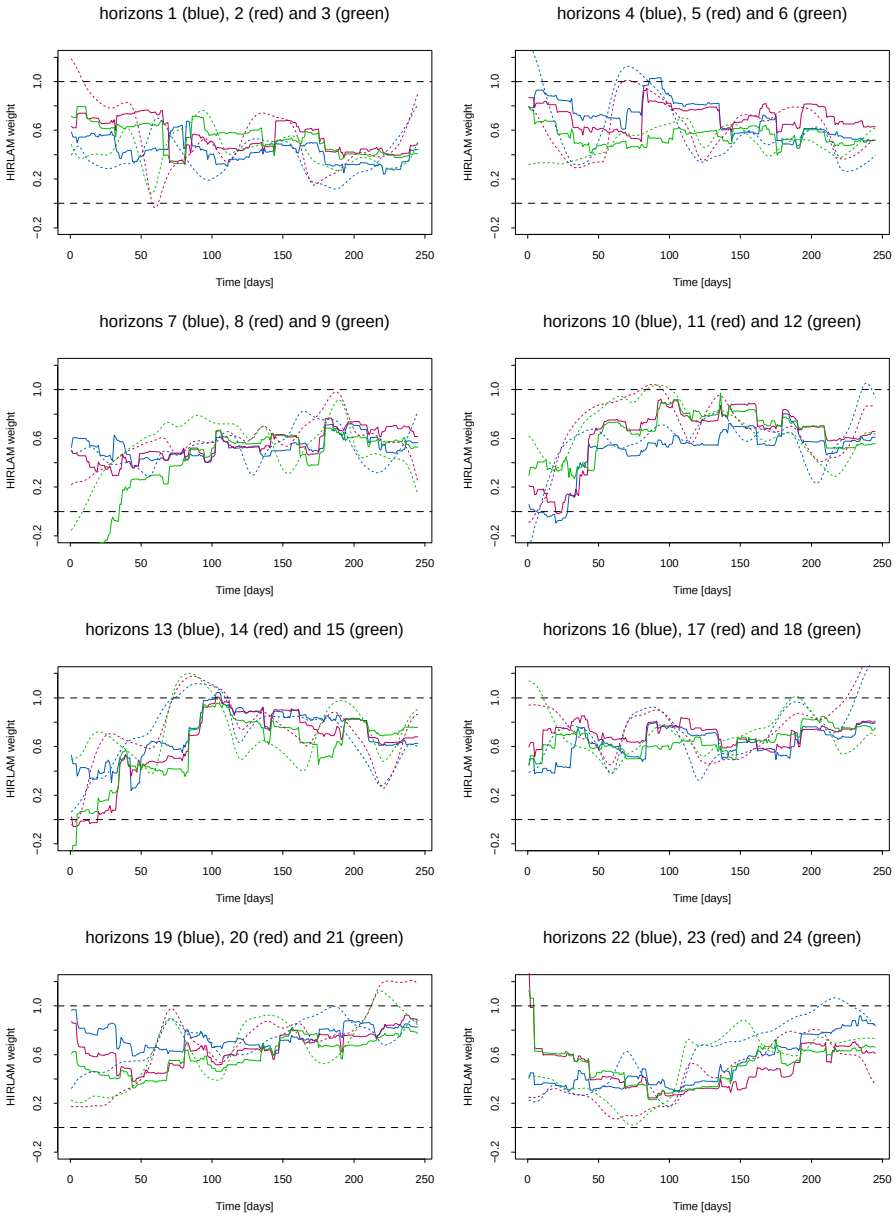


Figure C.3: Local fit (dashed lines) for HIRLAM weights in HIRLAM/MM5, compared with RLS estimations (solid lines) for corresponding combination.

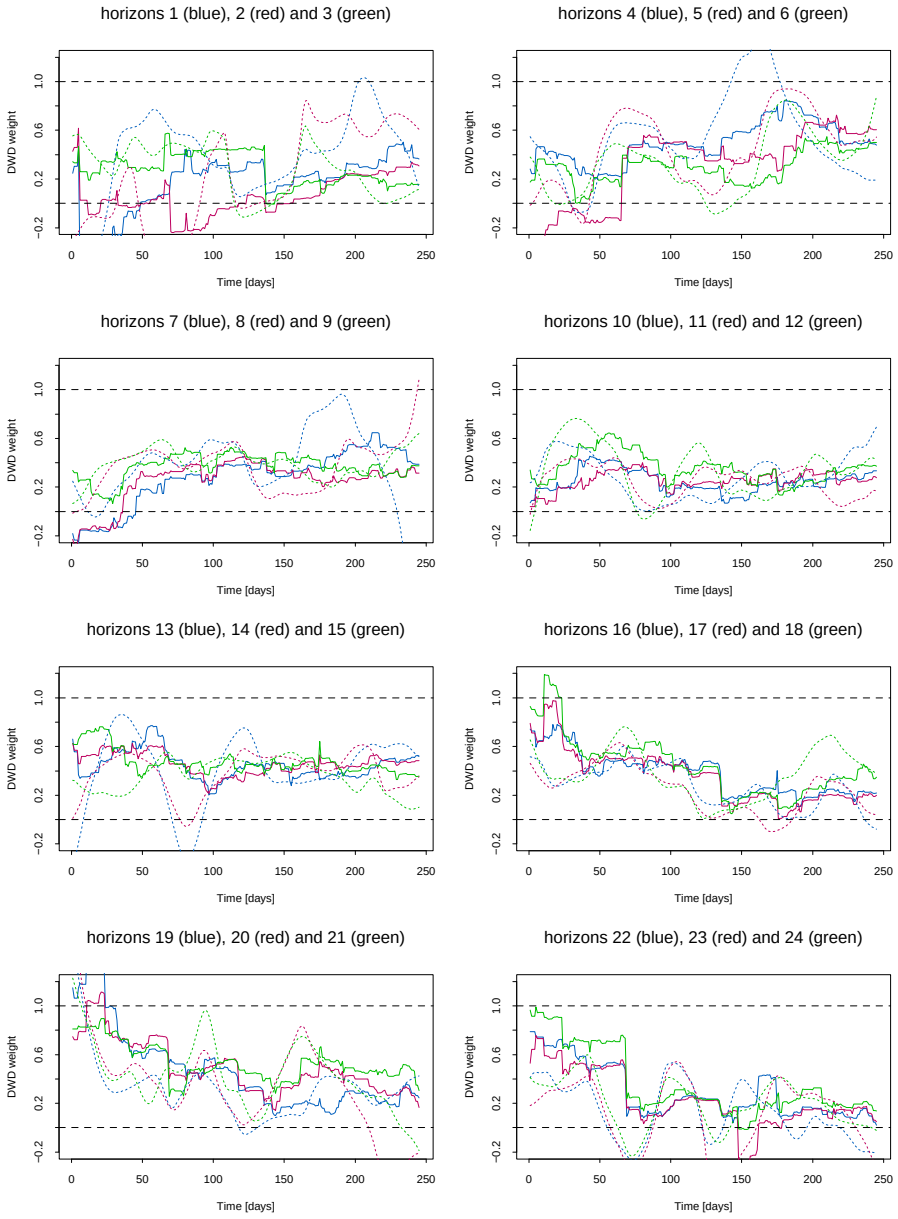


Figure C.4: Local fit (dashed lines) for DWD weights in DWD/HIRLAM/MM5, compared with RLS estimations (solid lines) for corresponding combination.

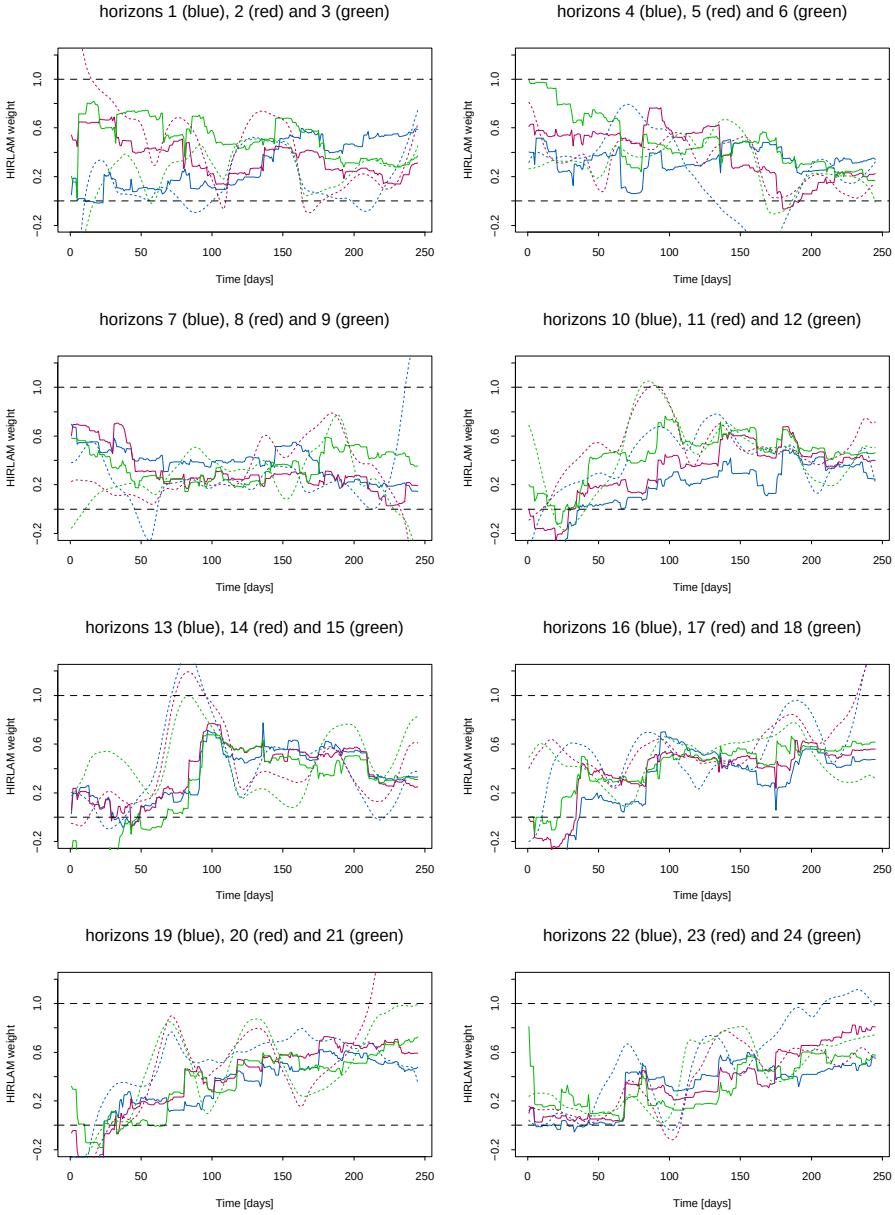


Figure C.5: Local fit (dashed lines) for HIRLAM weights in DWD/HIRLAM/MM5, compared with RLS estimations (solid lines) for corresponding combination.

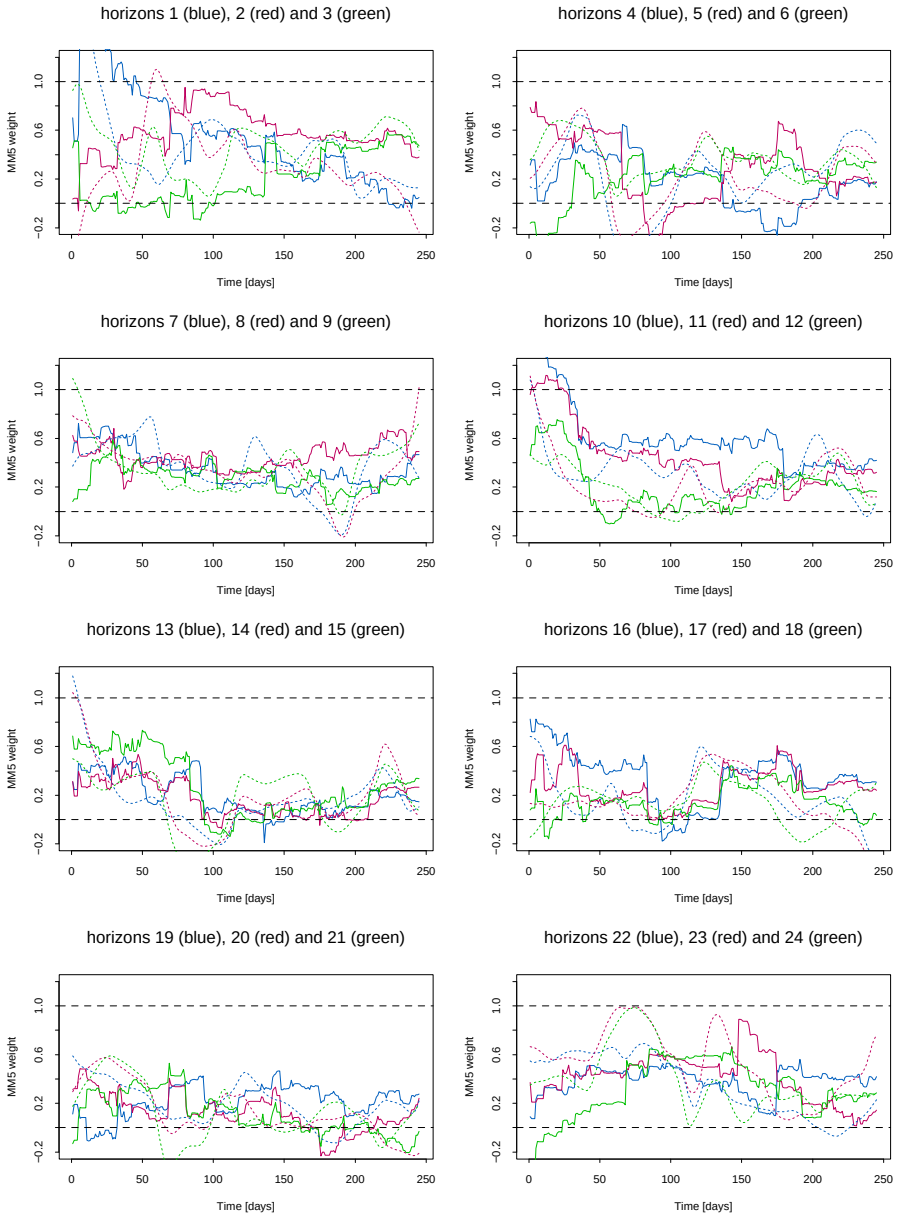


Figure C.6: Local fit (dashed lines) for MM5 weights in DWD/HIRLAM/MM5, compared with RLS estimations (solid lines) for corresponding combination.

Appendix D

Coplots for MET forecasts to estimate weights

The appendix includes the coplots where the MET forecasts are used to interpret the weights from the local fit. Only plots for DWD weight in DWD/HIRLAM are considered.

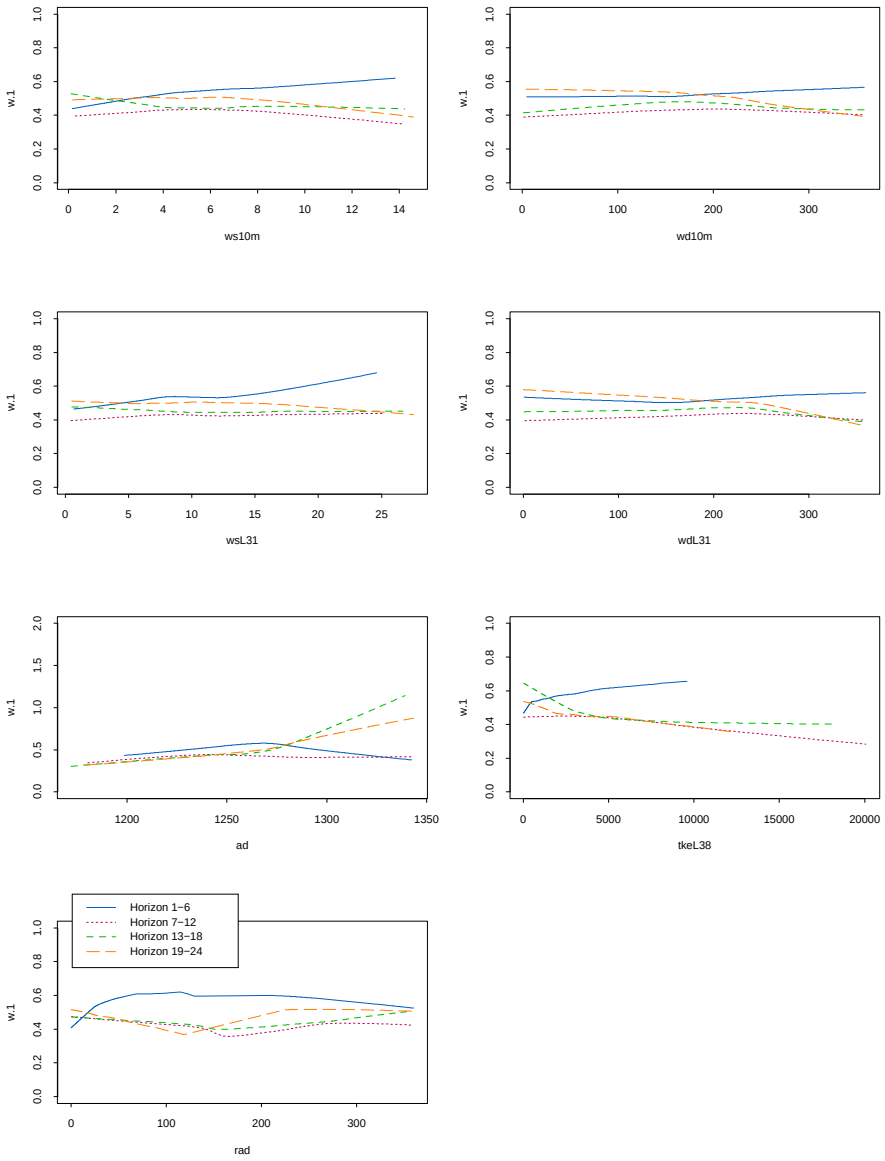


Figure D.1: Coplots with **horizon** variable partitioned.

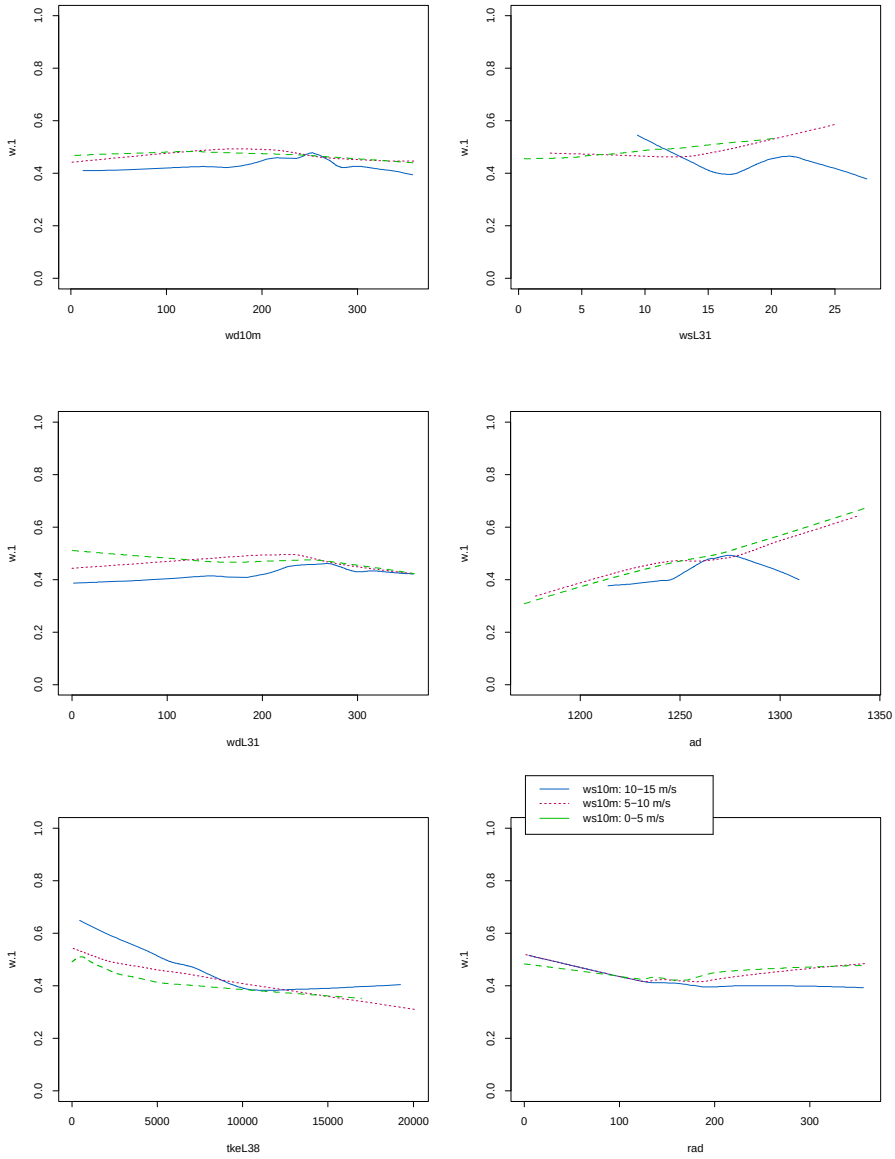


Figure D.2: Coplots with **ws10m** variable partitioned.

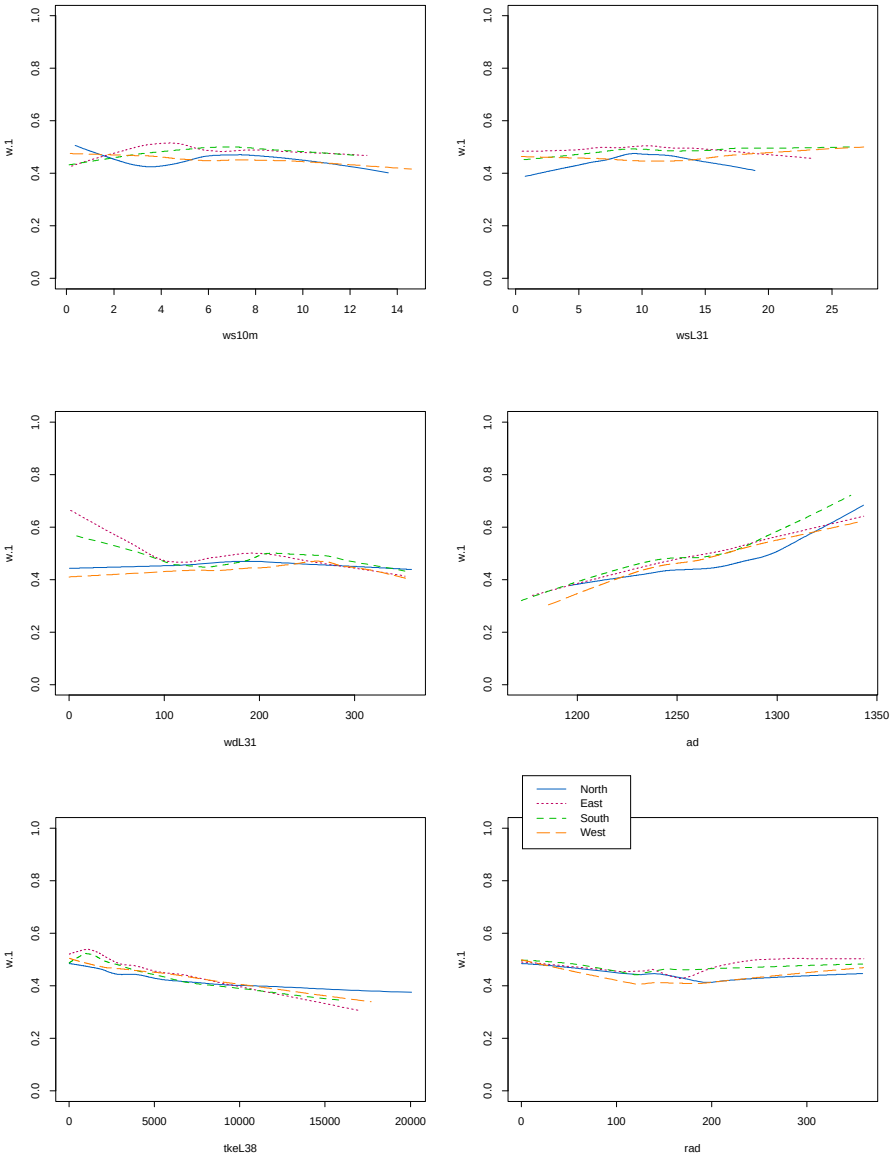


Figure D.3: Coplots with **wd10m** variable partitioned.

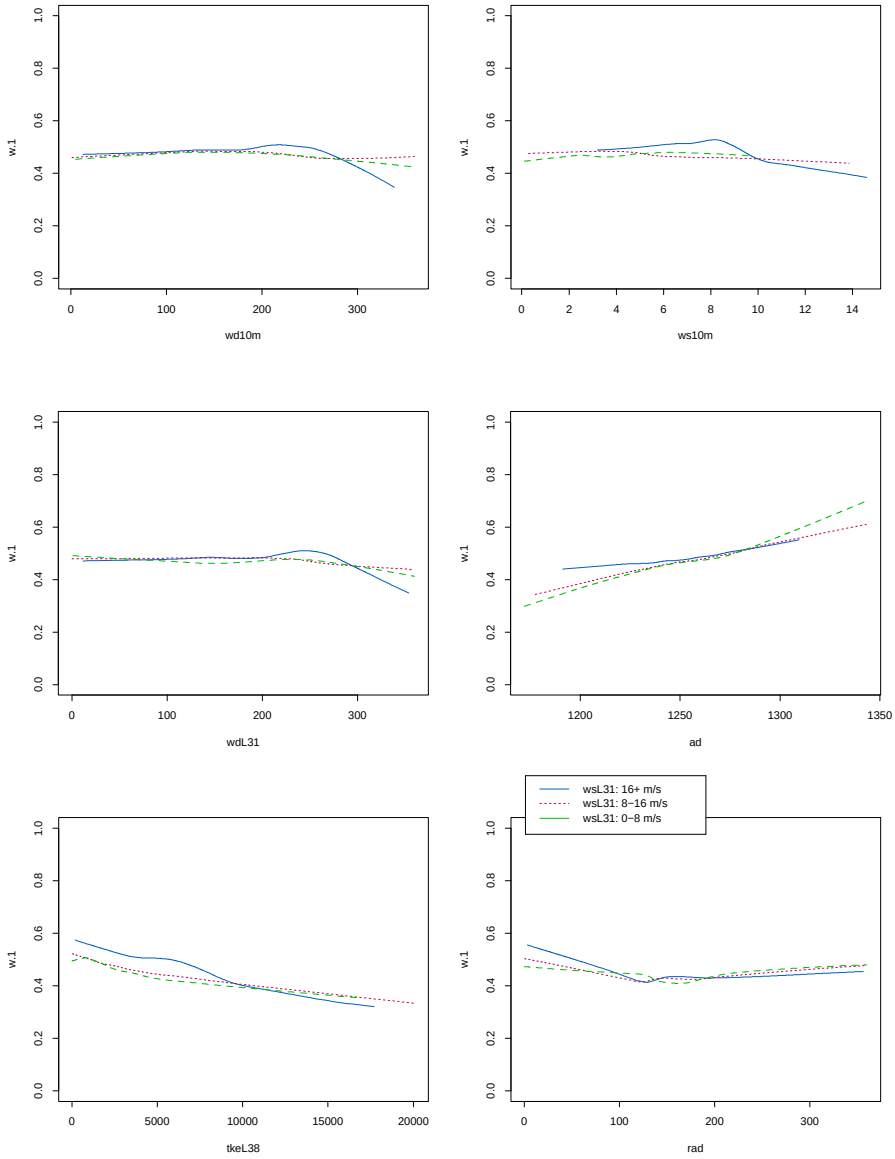


Figure D.4: Coplots with **wsL31** variable partitioned.

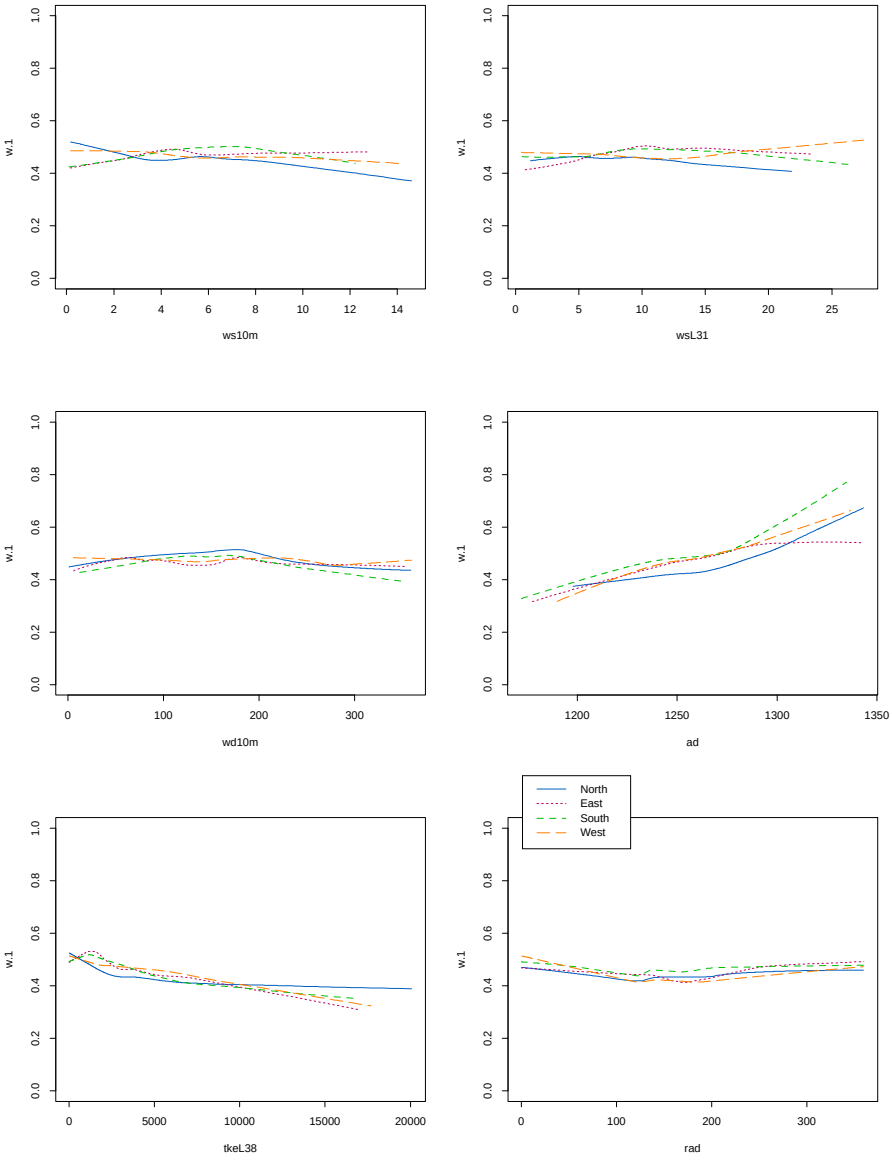


Figure D.5: Coplots with **wdL31** variable partitioned.

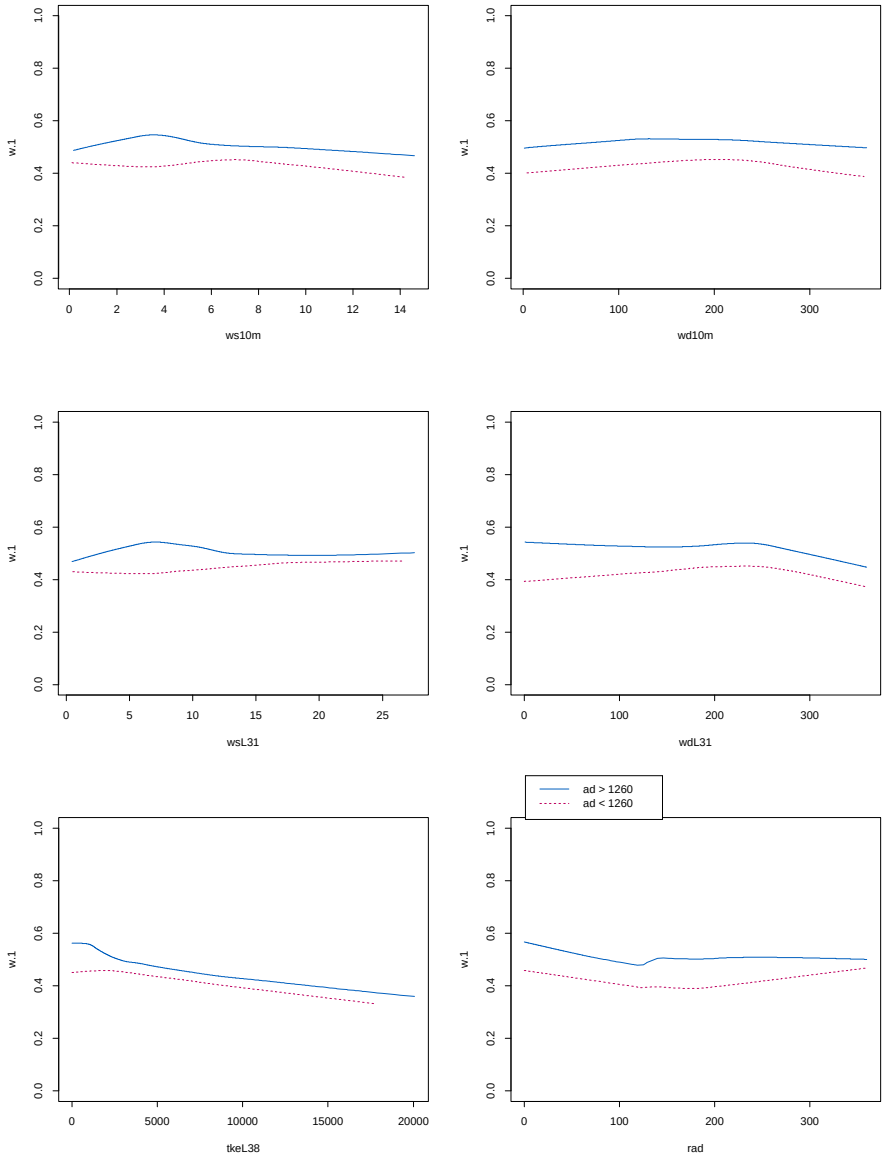


Figure D.6: Coplots with **ad** variable partitioned.

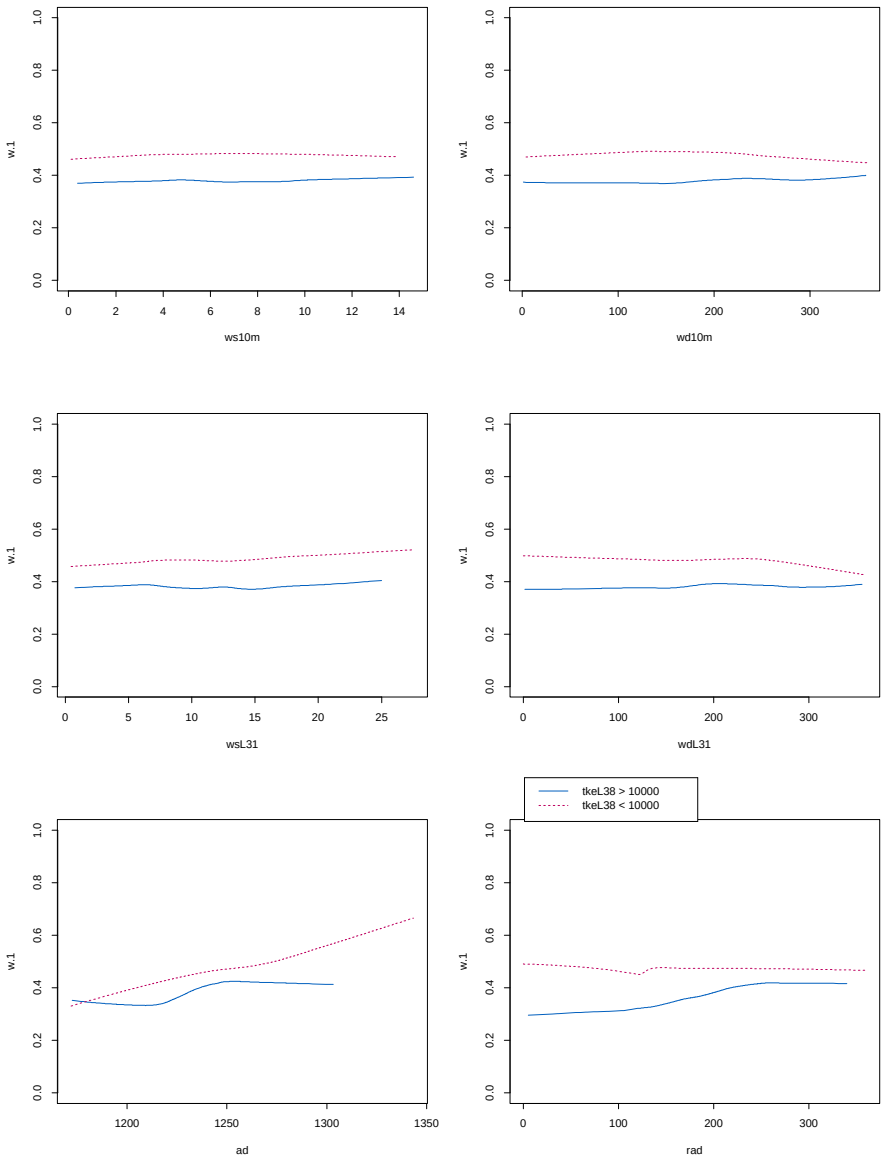


Figure D.7: Coplots with **tkeL38** variable partitioned.

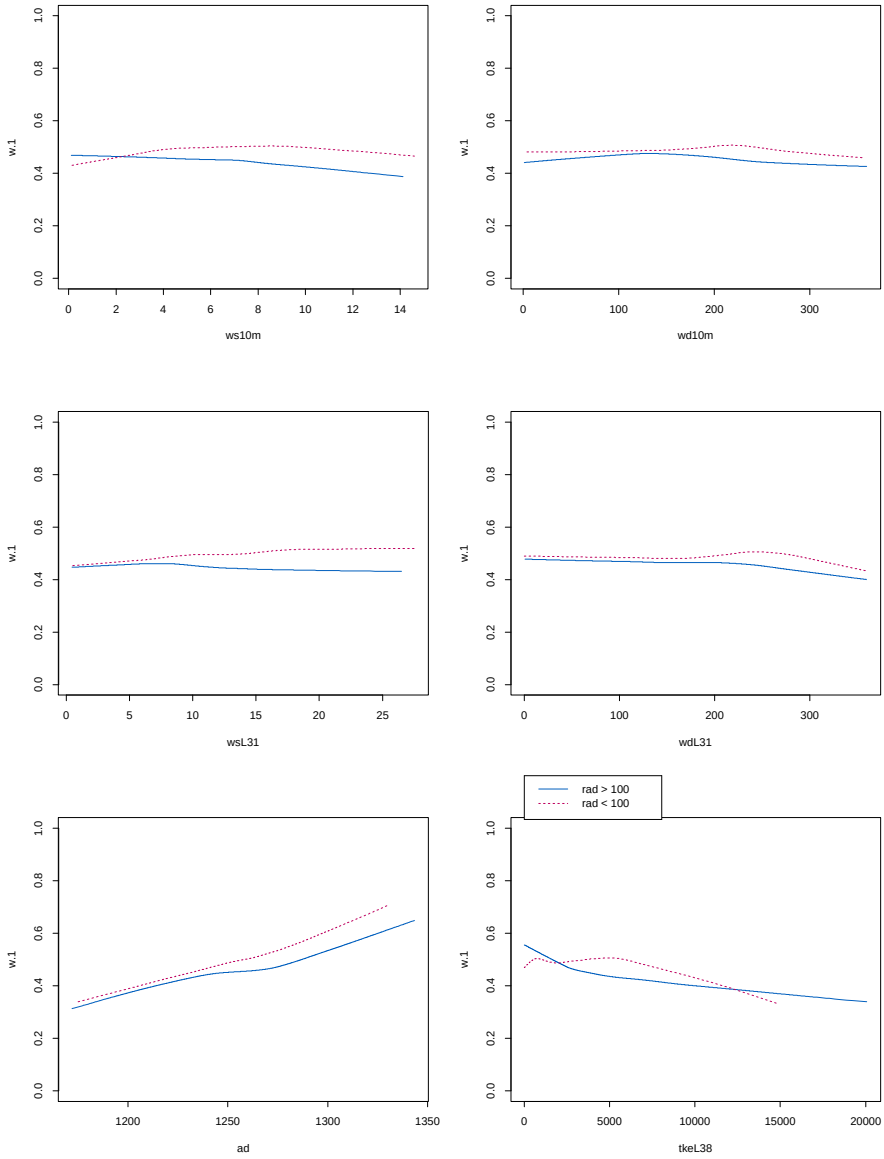


Figure D.8: Coplots with **rad** variable partitioned.

Bibliography

[1] D. W. Bunn. A bayesian approach to the linear combination of forecasts. *Operational Research Quarterly (1970-1977)*, 26(2):325–329, June 1975.

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

[3] W. S. Cleveland. *Multivariate Analysis and Its Applications*, volume 24 of *IMS Lecture Note-Monograph Series*, chapter Coplots, nonparametric regression, and conditionally parametric fits, pages 21–36. 1994.

[4] L. M. de Menezes and D. W. Bunn. The persistence of specification problems in the distribution of combined forecast errors. *International Journal of Forecasting*, 14(3):Pages 415–426, September 1998.

[5] C. W. J. Granger. Invited review: combining forecasts - twenty years later. pages 411–419, 2001.

[6] C. W. J. Granger and R. Ramananthan. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204, 1084.

[7] S. I. Gunter. Nonnegativity restricted least squares combinations. *International Journal of Forecasting*, 8:45–59, 1992.

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

[9] T. Hastie and R. Tibshirani. Varying-coefficient models, 1993.

[10] C. W. J. Granger J. M. Bates. The combination of forecasts. *Operational research quarterly*, 20(4):451–468, 1969.

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

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

[13] H. Madsen, H. A. Nielsen, and T. S. Nielsen. A tool for predicting the wind power production of off-shore wind plants. In *Proceedings of the Copenhagen Offshore Wind Conference & Exhibition*, Copenhagen, 2005. http://www.windpower.org/en/core.htm.

[14] H. A. Nielsen. *Parametric and Non-Parametric System Modelling*. PhD thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 1999.

[15] R. L. Winkler R. T. Clemen. Aggregating point estimates: A flexible modeling approach. *Management Science*, 39(4):501–516, Apr 1993.

[16] J. W. Taylor and D. W. Bunn. Investigating improvements in tha accuracy of prediction intervals for combinations of forecasts: A simulation study. *International Journal of Forecasting*, 15:325–339, 1999.