Improving solar radiation forecasts using advanced statistical post-processing methods

Kilian Bakker

Supervisors:

Maurice Schmeits, Kirien Whan, Wouter Knap (KNMI)

Jason Frank (Utrecht University)

Dated: January 2, 2019

Abstract

The increased usage of solar energy places additional importance on forecasts of solar radiation. Solar panel power production is primarily driven by the amount of solar radiation and it is therefore important to have accurate forecasts of solar radiation. Accurate forecasts that also give information on the forecast uncertainties can help users of solar energy make better solar radiation based decisions related to stability of the electrical grid. To achieve accurate forecasts of global radiation together with information about the uncertainty, we apply statistical post-processing techniques to the deterministic forecast from the Numerical Weather Prediction model HARMONIE, which runs at KNMI. We use regression based methods that determine relationships between observations of global radiation (made within the Netherlands network of automatic weather stations) and forecasts of various meteorological variables from HARMONIE. Those relationships are used to produce probabilistic forecasts of global radiation. We compare parametric methods that make assumptions on the distribution of the global radiation and non-parametric methods without any assumptions on the distribution. We find that both types of methods are able to generate probabilistic forecasts that improve the raw global radiation forecast from HARMONIE according to the root mean squared error (on the median) and the potential economic value. We also compare the different regression methods using various scoring metrics like the continuous ranked probability skill score, the Brier skill score and reliability diagrams. We find that quantile regression and generalized random forests generally perform best and the artificial neural network worst. Additionally we show that the non-parametric approaches use information from all predictors, whereas the parametric approaches look only at a limited number of predictors.

Contents

1 Introduction

Forecasts of meteorological variables are produced by projecting an initial state of the atmosphere into the future. This is done by applying physical evolution equations to the state of the atmosphere. The initial state of the atmosphere is calculated as precisely as possible with data assimilation methods that use observations and a previous forecast as inputs. Due to observational scarcity, errors in the previous forecast and imperfections in the data assimilation methods there are unavoidable errors in the initial state of the atmosphere. This makes the forecast also non-optimal. Since the atmosphere is a chaotic system, small errors in the initial state can lead to large errors in the final forecast. It can therefore happen that we forecast a different weather scenario compared to reality. There are more sources of errors in the forecast, for example incomplete knowledge of the physical equations describing the subgrid scale processes in the atmosphere, as well as discretization errors in both the initial state (in space) and the forecasts (in space and time). Although we cannot produce perfect forecasts, we can quantify the uncertainty in the forecasts. This can be done by generating an ensemble of slightly perturbed initial states and computing the forecasts from these different initial states. These forecasts together form a probabilistic forecast, giving a range of values that the forecast can take. We can also generate the probabilistic forecast by statistical post-processing techniques, even if we use only the forecast from the unperturbed initial state, called the deterministic forecast. Then forecasts of relevant meteorological variables are used as predictors to improve the forecast of a specific variable and quantify the uncertainty in it (i.e. making the forecast probabilistic). The probabilistic forecast gives much more information about the meteorological variable we are forecasting and is therefore of greater value.

The meteorological variable we study in this thesis is the total amount of solar radiation reaching the Earth's surface, called the global radiation. We apply statistical post-processing techniques to make probabilistic forecasts of the global radiation. This can be very beneficial for users of solar energy (the part of society generating solar power and distributing it over the electricity grid). The power production of solar panels is directly related to the solar radiation that reaches the panels. Therefore with skillful and reliable probabilistic forecasts, users are better informed about the power production of the panels and can make better decisions based on this information (e.g. buying alternative sources of energy).

Quite some work has already been done on generating probabilistic forecasts of global radiation. For example Lorenz et al. (2009) produced 95% confidence intervals on the clear sky index (the global radiation divided by the radiation that would have been measured in the absence of clouds), meaning that in 95% of the cases the clear sky index falls in the confidence interval. The authors produced these confidence intervals by fitting a polynomial function to the standard deviation of the forecast errors and then took the appropriate values from the polynomial for the bounds of the confidence interval. The width of the confidence interval is a measure of the uncertainty in the forecasts and is therefore valuable for users. Instead of only predicting two quantiles that define the confidence interval, Verzijlbergh et al. (2015) forecast the complete distribution of the clear sky index. They first discretized the clear sky index into a finite number of bins and then fitted a normal distribution on the clear sky index for each bin separately. This normal distribution is fitted based on relevant meteorological variables, meaning that the mean and spread of the distribution depend on relevant meteorological variables. The spread in the distribution then gives information about the uncertainty. Other distributions were also used for predicting the radiation, such as the beta distribution in Fatemi et al. (2018) and the gamma distribution in Bracale et al. (2013).

Massidda and Marrocu (2018) produced probabilistic forecasts in a different way. Instead of fitting some chosen distribution, they fitted only the quantiles of the distribution, a technique called quantile regression. They also extended this to a method using regression trees, namely the gradient boosted regression trees method. These methods both fit the quantiles of the distribution and the distance between different quantiles provides information about the uncertainty. Another method using regression trees is the quantile regression forests method, used in Almeida et al. (2015). Cervone et al. (2017) used the neural networks method, which is currently a popular method.

There are also various reviews giving an overview of papers using regression methods to produce probabilistic forecasts of solar radiation or solar power production. For example Antonanzas et al. (2016) reviewed artificial neural networks, random forests, support vector machines and k-nearest neighbours. Van der Meer et al. (2018) considered quantile regression, quantile regression forests and gradient boosting and Voyant et al. (2017) reviewed artificial neural networks, support vector machines, random forests, k-nearest neighbours and Markov chains. They also compared supervised and unsupervised learning of the methods.

There are also some papers comparing multiple methods on the same data set to see which methods perform best on their data set. They produced probabilistic forecasts for the different methods and compared them with some verification measure. For example David et al. (2018) made a comparison between probabilistic forecasts of solar radiation produced by random forests, neural networks, (weighted) quantile regression and gradient boosting methods. They used only past observations of the global radiation and the clear sky index as predictors. They came to the conclusion that (weighted) quantile regression and gradient boosted regression trees performed best on their data set, but the rest of the methods was also close to the best performing one. Mohammed and Aung (2016) made a comparison between a number of methods, this time focusing on probabilistic forecasts of solar power production. They used Numerical weather prediction (NWP) output from the European Centre for Medium-Range Weather Forecasts (ECMWF) as predictors for producing probabilistic forecasts. They compared a number of methods including random forests, neural networks, support vector machines, k-nearest neighbours, gradient boosting, lasso and ridge regression. They found that k-nearest neighbours works the best on their data set.

This last approach of comparing multiple regression methods using NWP output is also what we do in this thesis. We apply both parametric methods (where we make assumptions on the distribution of global radiation) and non-parametric methods (where we make no assumptions on the distribution). The parametric methods consist of fitting a gamma and a truncated normal distribution (Rigby and Stasinopoulos (2005)) and on fitting quantiles of the radiation using quantile regression (Koenker (2005)). The non-parametric methods consist of the quantile regression forests (Meinshausen (2006)), generalized random forests (Athey et al. (2016)), gradient boosted regression trees (Friedman (2001) and Friedman (2002)) and artificial neural networks (Cannon (2011) and Cannon (2018a)). All of these methods use forecasts of various meteorological variables as predictors. For this thesis we take temperature, humidity, cloud, radiation and aerosol forecasts as predictors. These predictors come mostly from the high-resolution NWP model HARMONIE, which runs at KNMI. Also some predictors come from the NWP model CAMS, which runs at ECMWF. Next to predictors from NWP models, we add space and time related predictors, such as the day of the year. For the predictand we use hourly global radiation observations from the Netherlands network of automatic weather stations. This consists of measurements of global radiation at 30 stations in the Netherlands.
The methods produce probabilistic forecasts of the global radiation using relationships between the predictors and predictand. The probabilistic forecasts are verified with multiple scoring metrics to see which method performs best. These include the continuous ranked probability skill score, Brier skill score and reliability diagrams (e.g. Wilks (2011)). Additionally we also show the improvement in accuracy of the median of the probabilistic forecasts with respect to the raw forecast (the global radiation forecast from HARMONIE). This is shown in terms of the mean absolute error and the root mean squared error. The improvement of the full probabilistic forecasts with respect to the raw forecast is shown in terms of potential economic value (e.g. Richardson (2000)). The potential economic value is an economically relevant and objective way of comparing raw forecasts with the probabilistic forecasts generated by the methods.

The thesis is structured as follows: in section 2 we mathematically describe all the methods and formulate the scoring metrics that we use. In section 3 we describe our research setting, including information about the data and the research procedure that we apply for generating the results. The results are presented in section 4 and the thesis is finished with a conclusion in section 5.

2 Theoretical background

2.1 Statistical post-processing

Statistical post-processing techniques are used to improve forecasts of meteorological variables, such as solar radiation, by means of removing systematic biases and by making a probabilistic forecast. This quantifies the uncertainty in the forecast of the meteorological variable. This is accomplished by looking at the effects that other meteorological variables (e.g. temperature, humidity or cloud cover) have on the variable we want to forecast. We take the observations for the variable we are forecasting as the predictand and the forecasts for the other meteorological variables as the predictors. Because we want to improve the already existing forecast, called the raw forecast, we also take this raw forecast as one of the predictors. We use observations for the variable of interest and therefore we look at some past time for which there are observations available to find the relationships between the predictors and the predictand. Then by using the values for the predictors together with the relationships found, we produce new forecasts of the predictand. When the relationships are accurate, we are generally improving the raw forecast through the removal of biases and by quantifying the uncertainty in the forecast. There are two different approaches to statistical modelling, parametric and non-parametric, and these are explained in detail in the next sections.

2.2 Parametric methods

For the parametric approach we make assumptions about the relationships between the variables. For producing probabilistic forecasts we also assume that the predictand follows a given distribution. These assumptions can be expressed in terms of a set of parameters and the approach is to optimize these parameters by different procedures where we fit a relationship between the predictand and the predictors. For deterministic forecasts the simplest case of the fitting procedure is linear regression, where the relationship between the predictand and the predictors is assumed to be linear. This is expressed with the following linear regression equation:

$$O = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{1}$$

where O is the predictand, which is the observation of the variable we are interested in. The $X_j, j = 1, ..., p$, are the predictors, consisting of forecasts of other meteorological variables. The β_0 is a scalar constant called the intercept or regression constant and the β_j, $j = 1, ..., p$, are scalar constants called the regression parameters. Denote them together by the vector $\vec{\beta} = \{\beta_0, \beta_1, ..., \beta_p\}$. Suppose now that we have n values for the predictand and each predictor and we want to find the best linear relationship between the n values for the predictand and predictors. For that we minimize the sum of squared residuals SSR over $\vec{\beta}$, defined by:

$$SSR(\vec{\beta}) = \sum_{i=1}^{n} \epsilon_i^2, \tag{2}$$

$$\epsilon_i = O_i - (X\vec{\beta})_i, i = 1, ..., n \tag{3}$$

where $\{O_i, i = 1, ..., n\}$ are the values for the predictand and X is the matrix consisting of the values for the predictors:

$$X = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix},$$

where $X_{i,j}$ denotes the i-th value of the j-th predictor for $i = 1, ..., n$ and $j = 1, ..., p$.
This leads to the best (linear) fit between the predictand and all predictors. However, leaving out one or more predictors in Equation (1) can lead to a better fit, because using too many predictors can lead to overfitting (which means that we are fitting to noise instead of to the actual relationships between the predictors and predictand). To avoid overfitting we only use a subset of the set of predictors in Equation (1). The subset that gives the best fit between the predictors and predictand is found by a stepwise procedure, consisting of forward and backward steps. The procedure starts with only the intercept β_0 in Equation (1) and then minimizing the SSR over β_0 to find the optimal fit in this case. Then we calculate the Akaike Information Criterium (AIC) over the fit, which is defined as:

$$\text{AIC} = -2L(\vec{\beta}) + 2(\tilde{p} + 1), \tag{4}$$

where $\tilde{p}$ is the number of predictors in the fit, which is 0 in this case. $L(\vec{\beta})$ is the log-likelihood function and we make the assumption that the residuals are independent and normally distributed with mean 0 and constant variance σ^2. We can estimate this variance with the sampling variance:

$$\hat{\sigma}^2 = \frac{1}{n - \tilde{p} - 1} \sum_{i=1}^{n} \epsilon_i^2,$$

We then have for the log-likelihood function:

$$\begin{aligned} L(\vec{\beta}) &= \log\left(\prod_{i=1}^{n} \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{\frac{1}{2\hat{\sigma}^2}\epsilon_i^2}\right) \\ &= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\epsilon_i^2 \end{aligned}$$

Inserting this in Equation (4) gives the formula for the AIC that we use. The forward procedure starts by adding one of the predictors to Equation (1) and minimizing the SSR and then calculating the AIC over the fit. This is then repeated for all of the predictors, which leads to an AIC value for each predictor. An increase in the likelihood function indicates a better fit. This is the same as a decrease in the log-likelihood and consequently also in the AIC. Therefore the predictor with the lowest AIC value leads to the best fit in comparison to the rest of the predictors. We also compare this lowest AIC value with the AIC of the fit with only the intercept, and if it is lower then we keep the predictor in Equation (1). Then we continue by adding a second predictor by looking among all remaining predictors which one leads (in combination with the predictor that is already added into the model) to the lowest AIC value. If this lowest AIC value is also lower than the AIC of the fit with only the first predictor, then we add this second predictor to the model. In the same way we also try to add a third, fourth, etc. predictor to the model. The forward procedure is iterated until we are not able to lower the AIC anymore.

At the same time that the forward procedure happens, we also apply the backward procedure. Each time we add a predictor to the model, we also look among all predictors that are already in the model if we can take one of them out of the model to achieve a better fit (i.e. a lower value for the AIC). If this is the case then we take the predictor out to achieve the better fit and else we keep all the predictors. taking one of the predictors out of the model can lead to a better fit, because it can happen that we overfit and taking one predictor out turns out to solve that (partly). The goal of the forward procedure is therefore to find the best fit and the backward procedure tries to avoid overfitting. To make the chance for overfitting even lower, we can also decide to allow only a certain number of predictors in the fit. Once this maximum number of predictors is reached, the procedure stops and we calculate our final fit by minimizing the SSR. Then we use the final fit to make new predictions.

These predictions are deterministic and we can extend the procedure described above to form probabilistic predictions. There are two approaches we can follow. The first one is that we assume some distribution to the observations and fit the parameters that define the distribution with a linear regression procedure. Then we draw quantiles from the fitted distribution after the fitting has finished. This is done by denoting the fitted cumulative distribution function (CDF) by $G_{\tilde{O}}$ and then for each quantile q calculating $\inf\{O|G_{\tilde{O}}(O) \geq q\}$ to find the value of the observations that corresponds to the q-th quantile of the distribution function $G_{\tilde{O}}$.

The other approach is to directly fit the quantiles from the distribution by assuming some formula with parameters for each quantile and fit the parameters in this formula. The advantage of the first approach is that we form the complete probability distribution, instead of only looking at some quantiles of the distribution. Also the second approach is done for each quantile separately, while the first approach is only done on the parameters of the distribution, and in general we take more quantiles than there are parameters in a distribution. The second approach has however the advantage that there is more freedom in the assumed distribution. Because we can take a different formula for each quantile, the distribution that belongs to that quantiles can take almost any form we would want.

For the first approach of fitting the complete distribution, suppose that the distribution of the observations is described by the parameters $\{\mu, \sigma, \tau, \nu\}$, which stand for the location, scale, skewness and kurtosis parameters. We start with fitting only an intercept term to the four parameters. Then we fit a model for μ on top of it by the linear regression procedure:

$$\mu = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p. \tag{5}$$

We minimize the SSR to find the optimal value for $\vec{\beta}$. We also minimize the AIC with the stepwise procedure to find the optimal predictors for in the model. Then this is repeated for σ to build a model for σ on top of the model for μ. This is then continued in the same way for τ and ν. Because the parameters are fitted separately, each one has different optimal predictors and different optimal values for $\vec{\beta}$. After this procedure we have fits for all the parameters of the distribution and we can make probabilistic predictions by using the fitted model. We apply this approach in the programming language R, using the gamlss package (Rigby and Stasinopoulos (2018) and Rigby and Stasinopoulos (2005)).

For the second approach of fitting quantiles, we assume the linear regression Equation (1) for each quantile that we want to fit. Then, according to Koenker (2005), we are not minimizing the sum of squared residuals for all quantiles, but for each quantile q we minimize the weighted sum of (absolute) residuals WSR, where the weights depend on q:

$$WSR = \sum_{i=1}^{n} \rho_q(\epsilon_i) \tag{6}$$

$$\rho_q(\epsilon_i) = (q - \mathbb{1}_{\epsilon_i<0})\epsilon_i \tag{7}$$

Here $\mathbb{1}$ denotes the indicator function and ϵ_i is defined in Equation (3).

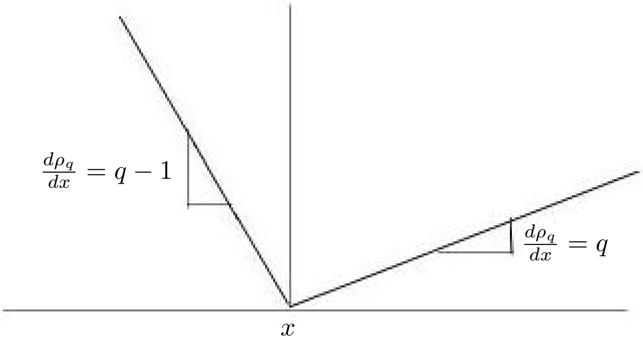


Figure 1: The function $\rho_q(\epsilon_i)$, which is called the quantile loss function.

To see that this indeed leads to a good fit for the q-th quantile, denote by $G_{\tilde{O}}$ the (unknown) CDF of the observations. We show that, in the limit of an infinite number of observations, minimizing the WSR gives the exact q-th quantile. With this infinite number of observations, we are not minimizing a sum, but an expectation, and we have the following:

$$\begin{aligned}
\mathbb{E}(\rho_q(\epsilon)) &= \int_{-\infty}^{\infty} \rho_q(\epsilon)\, dG_{\tilde{O}}(O) \\
&= \int_{-\infty}^{\infty} (q - \mathbb{1}_{O-X\beta<0})(O - X\beta)\, dG_{\tilde{O}}(O) \\
&= (q-1)\int_{-\infty}^{X\beta} O - X\beta\, dG_{\tilde{O}}(O) + q\int_{X\beta}^{\infty} O - X\beta\, dG_{\tilde{O}}(O)
\end{aligned}$$

The value $\hat{\beta}$ that minimizes this expression over β is where the derivative is equal to 0:

$$\begin{aligned}
0 &= (1-q)(G_{\tilde{O}}(X\hat{\beta}) - G_{\tilde{O}}(-\infty)) - q(G_{\tilde{O}}(\infty) - G_{\tilde{O}}(X\hat{\beta})) \\
&= G_{\tilde{O}}(X\hat{\beta}) - q
\end{aligned}$$

Therefore $G_{\tilde{O}}(X\hat{\beta}) = q$, so $X\hat{\beta}$ satisfies $\inf\{O|G_{\tilde{O}}(O) \geq q\}$ and we have that $X\hat{\beta}$ is the q-th quantile of the observations. This is the case for an infinite number of observations, so with a finite number of

observations we have an approximation of the q-th quantile. This approximation gets closer to the actual q-th quantile when we increase the number of observations.
In this approach the WSR is used for finding the optimal value for $\vec{\beta}$ and this applies to each quantile separately. Then for finding the optimal predictors we combine the fits for all the quantiles and calculate the average over the AIC values from all the fits. This average AIC value is minimized with the stepwise procedure. This leads to the optimal predictors for in the model, which are the same predictors for all the quantiles. Therefore for each quantile we have the same model set up, which gives a fair comparison between the different quantiles. By generating fits for a number of quantiles, we can form probabilistic predictions. The approach just described is applied in R using the quantreg package (Koenker (2018)).

2.3 Non-parametric methods

In the case of non-parametric regression methods we do not make assumptions on the relationships between the predictand and the predictors. This means that there are no parameters involved that we want to estimate. Non-parametric methods are also called machine learning methods, because the machine (the algorithm) uses the data to learn about (or find) the best relationships between the predictand and predictors. The learning procedure for finding the best relationships can be done in various ways and we discuss four of them in the coming sections. Most of the methods we discuss make use of binary regression trees, so we first explain how to construct these trees. The procedure is visualized in Figure 2.

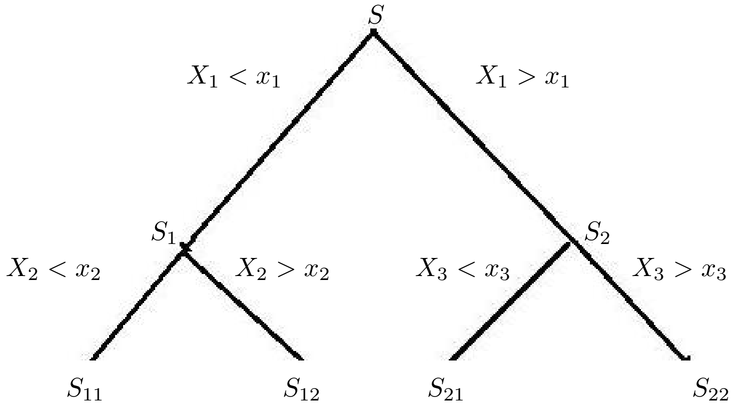


Figure 2: Example of a binary tree

The procedure starts with the set of predictors and predictand $S = \{O_i, X_{i,j} | i = 1, ..., n, j = 1, ..., p\}$. The set S is split into two subsets S_1 and S_2 according to a threshold T on one of the predictors, say X_k. Define $I = \{i | X_{i,k} < T\}$, then we have $S_1 = \{O_i, X_{i,j} | i \in I, j = 1, ..., p\}$ and $S_2 = \{O_i, X_{i,j} | i \notin I, j = 1, ..., p\}$. For the choice of T and k we maximize the homogeneity over all possible values for T and k, where the homogeneity H is defined by:

$$H(S, S_1, S_2) = |S|\text{var}(\{O_i | O_i \in S, i = 1, ..., n\}) - |S_1|\text{var}(\{O_i | i \in I\}) - |S_2|\text{var}(\{O_i | i \notin I\}), \quad (8)$$

where the size $|S|$ of a set S is defined as the number of observations in S.
For all possible values of k we go over all predictors. For all possible values for the threshold T, we only have to go over $n-1$ thresholds, where n is the number of values we have for the predictor X_k. By first ordering the values for the predictor, each of the $n-1$ thresholds is in between two consecutive values of the predictor out of the total n values.
Maximizing the homogeneity H means we maximize the decrease in variance of the observations between S and S_1, S_2. Doing so leads to subsets S_1 and S_2 which have lower variance in the observations then the original set S. Therefore the observations are closer to each other and can be described better in terms of the other predictors. We repeat this procedure a number of times. Suppose for example that we are in step s and we want to split the set $S_s = \{O_i, X_{i,j} | i \in I_{s-1}, j = 1, ..., p\}$ in $S_{s1} = \{O_i, X_{i,j} | i \in I_s, j = 1, ..., p\}$ and $S_{s2} = \{O_i, X_{i,j} | i \notin I_s, j = 1, ..., p\}$, where $I_s = \{i | i \in I_{s-1}, X_{i,k_s} < T_s\}$. Here the predictor X_{k_s} and the threshold T_s are found by maximizing the homogeneity $H(S_s, S_{s1}, S_{s2})$. By repeating the procedure

we keep reducing the variance of the observations in the resulting subsets. We stop the splitting procedure when we reach some stopping criterion, which can be that the homogeneity is not above some chosen constant value. This means that the sum of the variances in the subsets has not decreased enough with respect to the variance of the set that is the union of the subsets. We can also stop the procedure when the size of the set has gone below some chosen constant value, meaning that there are not enough observations O_i left in the set to make another split on the set.

This procedure creates a binary tree with S at the top and the splits as branches leading to the subsets, as visualized in Figure 2. If we have a new set of predictors S_{new}, then we make a new prediction by following the branches in the tree that correspond to the values of the predictors being above or below the thresholds until we reach the terminal node L. This terminal node L always exists and is unique, because there is exactly one path of branches that leads to L. The deterministic prediction is then the mean of the observations in L and the probabilistic prediction is produced by taking quantiles from the observations in L.

The binary regression tree is formed only on one set of data and can therefore easily overfit on the data. This can partly be solved by stopping the formation of the tree at an earlier stage. Another approach that is better in reducing the chance for overfitting, is to form multiple trees on different parts of the data and combining the results. This last approach will be used in the methods described in the two coming sections.

2.3.1 Random Forests

The random forest method grows many binary regression trees and aggregates them to generate predictions. We follow the approach described in Breiman (2001) and Meinshausen (2006), which grows a certain number nt of trees: $TR_1, ...TR_{nt}$. For a new set of predictors S_{new}, there is a unique terminal node L_k in the tree TR_k for each $k = 1, ..., nt$. To form the deterministic prediction $F_{S_{new}}$ we calculate the mean of the observations in all the terminal nodes weighted by the frequency of appearance in the terminal nodes:

$$F_{S_{new}} = \sum_{i=1}^{n} \frac{\sum\limits_{k=1}^{nt} \mathbb{1}_{O_i \in L_k}}{\sum\limits_{k=1}^{nt} |L_k|} O_i$$

For the probabilistic prediction, we not only look at the mean, but at the full distribution. Denote by $G_{\tilde{O}}$ the cumulative distribution function (CDF) of the observations. It holds that $G_{\tilde{O}}(O) = \mathbb{P}(\tilde{O} \leq O) = \mathbb{E}(\mathbb{1}_{\tilde{O} \leq O})$. We estimate this expectation by the mean: $\hat{G}_{\tilde{O}}(O) = \frac{1}{n} \sum\limits_{i=1}^{n} \mathbb{1}_{O_i \leq O}$. This is the case when all observations have equal weight for the CDF, but here we are dealing with multiple terminal nodes and we take the frequency of appearance in the terminal nodes as the weights for the observations:

$$\hat{\hat{G}}_{\tilde{O}}(O) = \sum_{i=1}^{n} \frac{\sum\limits_{k=1}^{nt} \mathbb{1}_{O_i \in L_k}}{\sum\limits_{k=1}^{nt} |L_k|} \mathbb{1}_{O_i \leq O}, \tag{9}$$

We can draw quantiles from the CDF $\hat{\hat{G}}$ by using that the value corresponding to the q-th quantile is found by calculating $\inf\{O | \hat{\hat{G}}_{\tilde{O}}(O) \geq q\}$. We then form the probabilistic prediction $F_{S_{new}}$ by applying the procedure just described to all the quantiles of interest. This probabilistic approach is called quantile regression forests.

The procedures for generating deterministic or probabilistic predictions still depend very much on the data we have and can therefore overfit on the data, causing bad predictions for a new data set. To minimize overfitting, each tree is grown on a different subset of the data. The different subsets are randomly generated using bootstrap samples.

To make the procedure even more independent of the data set, at each splitting point only part of the predictors is chosen to look over for making the best split. This part of predictors is randomly chosen by bootstrap samples. The predictors that are chosen changes for each split and that reduces the chance for overfitting.

For producing results for the quantile regression forests method we use the R package quantregForest (Meinshausen (2017)).

2.3.2 Generalized random forests

The generalized random forests (GRF) method is a extension of the random forests method and is described in Athey et al. (2016). It adjusts the procedure for making splits in growing trees. In the random forests method, trees are grown by splits that maximize the homogeneity, but in the GRF method we make splits in growing trees by maximizing the Δ function:

$$\Delta(S, S_1, S_2) = \frac{|S_1|}{|S|}\frac{|S_2|}{|S|}(F_{S_1} - F_{S_2})^2$$

Here F_{S_j} is the prediction made considering S_j to be the terminal node for $j = 1, 2$. For a deterministic prediction this is the mean of the observations in S_j and for a probabilistic prediction this is the prediction for one quantile of the observations in S_j. The Δ function is then an average over all quantiles.

Whereas the homogeneity by random forests is a function that only has terms depending on either S, S_1 or S_2, the Δ function has a term depending on S_1 and S_2 at the same time and can therefore be seen as some form of heterogeneity. Maximizing the Δ function means that we are maximizing the distance between the predictions in S_1 and S_2. This leads to the subsets S_1 and S_2 becoming more distinct from the rest, which should lead to better predictions in the end. To reduce the computational time, we use an approximation of the Δ function, as described in Athey et al. (2016).

For the GRF method we also construct a forest of nt binary trees by maximizing the Δ function for each split. Then we make predictions for new sets of predictors in the same way as is described in Section 2.3.1.

To reduce the chance for overfitting we use the same procedure as in the random forests method. Each tree is formed on a random subset of the data and for each split there is only a subset of the predictors available to split upon. For this method we use the R package grf (Tibshirani et al. (2018)).

2.3.3 Gradient boosted regression trees

The random forests and generalized random forests are bagging methods, meaning that they repeat some procedure (forming a binary tree) a number of times, each time starting from the original observations, and aggregating the results. A different approach is to use a boosting method. This means that the method repeats some procedure a number of times, but each time it starts with the predictions from the previous step. Then each new step tries to improve upon the previous step and the last step gives the best results. In the case of the gradient boosted regression trees method the procedure consists of forming a binary tree on the gradient of the errors (the direction that the errors are moving to) between the predictions and observations and using it to improve the predictions made in the previous step. This is done by looking at some error function E_F between the predictions F in the previous step and the actual observations O. In the case that we want a deterministic prediction, we take E_F equal to the squared error: $E_F(O, F) = (O - F)^2$. In the case of a probabilistic prediction, we apply the boosting method for each quantile q separately and take the error function to be the quantile loss function $E_F(O, F) = \rho_q(O - F)$ (where ρ_q is defined in Equation (7)). The quantiles are sorted at the end to prevent crossings between the quantiles.

For the tree based approach in the boosting framework, we follow the procedure described in Friedman (2001) and Friedman (2002). The starting predictions are calculated as $F_0(O_i) = \arg\min_p \sum_{i=1}^{n} E_F(O_i, p)$. Then we apply the iterative procedure of improving the predictions. Suppose we are in step m of the procedure, where we have predictions $F_m(O_i)$ for the observations $\{O_i, i = 1, ..., n\}$. We calculate the derivative of the error function between the predictions and observations and evaluate the derivative at the predictions $F_m(O_i)$:

$$g_m^i = -\left.\frac{\delta E_F(O_i, F)}{\delta F}\right|_{F=F_m(O_i)}, i = 1, ..., n$$

We make predictions for the negative gradients g_m^i using the predictors. This is done by forming a new binary tree, where we take the predictand set to be $\{g_m^i, i = 1, ...n\}$ and the predictor set remains the same as in the previous step. This means we fit the binary tree on the negative gradient of the errors. From this binary tree we can make predictions for the negative gradients. Suppose that the constructed tree has N terminal nodes: $L_{m,j}, j = 1, ..., N$. To update our predictions $F_m(O_i)$ of the observations we replace the negative gradients in the terminal nodes with their corresponding observations and calculate the following:

$$
\begin{aligned}
F_{m+1}(O_i) &= F_m(O_i) + v\sum_{j=1}^{N} p_{m,j}\mathbb{1}_{O_i \in L_{m,j}}, i = 1, ..., n, \\
p_{m,j} &= \arg\min_{p} \sum_{\substack{i=1, \\ O_i \in L_{m,j}}}^{n} E_F(O_i, F_m(O_i) + p),
\end{aligned}
$$

where v is a chosen constant called the learning rate. v is generally chosen to be small (≈ 0.1) to achieve better predictions. When the learning rate is too high, then we can boost the predictions too much, causing them to overshoot the minimum in the errors. With the appropriate learning rate, the predictions $F_m(O_i)$ are boosted towards the updated predictions $F_{m+1}(O_i)$. which are closer to the actual observations (i.e. they are closer to the minimum in the error function).

We apply the iterative procedure for a given number of iterations to boost the predictions towards the observations. The predictions are not guaranteed to converge towards the observations, but they get closer if we apply more iterations. We choose enough iterations to let the predictions become sufficiently close to the observations. To reduce the chance for overfitting, in each step of the procedure where we form a new binary tree, we form this tree only on a subset of the data. This subset is randomly chosen by bootstrap samples and is therefore different each time. Consequently, in each step we only update the predictions for the observations that are in the subset that was chosen for the current step.

The gradient boosted regression trees method is applied in R using the package gbm (Ridgeway (2018)).

2.3.4 Artificial neural networks

Another non-parametric approach is an artificial neural network. This method is described in Cannon (2011) and Cannon (2018a). It is a boosting method that boosts the predictions towards the observations, just like in the gradient boosted regression trees method. There it was done using binary trees, but in the neural networks method it is done using an input layer, one or more intermediate hidden layers and an output layer. In the hidden layers we adjust the input towards the output by multiplication with weights w_j, adding some bias b and applying an activation function f. We take this activation function to be the Sigmoid function $f(x) = \frac{1}{1+e^x}$. This means that if a hidden layer has inputs z_j for $j = 1, ..., J$, then the output in the hidden layer is:

$$
\text{Output} = f(\sum_{j=1}^{J} w_j z_j + b) \tag{10}
$$

An extra component with deep hidden layers can also added as is visualized in Figure 3. Both in the hidden layers and in the deep hidden layers are inputs adjusted towards the outputs. This gives a wide range of calculations that can be done in the (deep) hidden layers. The predictors form the input layer of the neural network and in the (deep) hidden layers these are transformed to one single number, the prediction F corresponding to the predictors. This is the output of the neural network. The desired output is the observation O corresponding to the predictors and we bring the prediction closer towards the observation in an iterative procedure by minimizing some error function over the weights and biases.

The first step in the procedure consists of plugging in the predictors for all $i = 1, ..., n$ and then applying the starting conditions in the hidden layers. The starting conditions for the weights are drawn from a uniform distribution on the interval $[-0.5, 0.5]$. The bias has a starting condition at zero. With these starting conditions we produce predictions $\{F_0(O_i), i = 1, ..., n\}$ as the output.

Suppose now that we are in step m of the procedure, where we have predictions $F_m(O_i)$ for $i = 1, ..., n$. The predictions are compared with the observations $\{O_i, i = 1, ..., n\}$ using an error function E_F. The error function is taken to be sum of squared errors $E_F(O, F) = \sum_{i=1}^{n}(O_i - F_i)^2$ for deterministic predictions and the sum over the quantile loss functions $E_F(O, F) = \sum_{i=1}^{n} \rho_q(O_i - F_i)$ for probabilistic predictions (where ρ_q is defined in Equation (7)).

The error function depends on the predictions and therefore also on the weights and biases and by back propagation, i.e. filling in the formulas applied in the hidden layers in E_F, we find an expression for the dependence. Next, the error function is minimized over the weights and biases. This is done with a

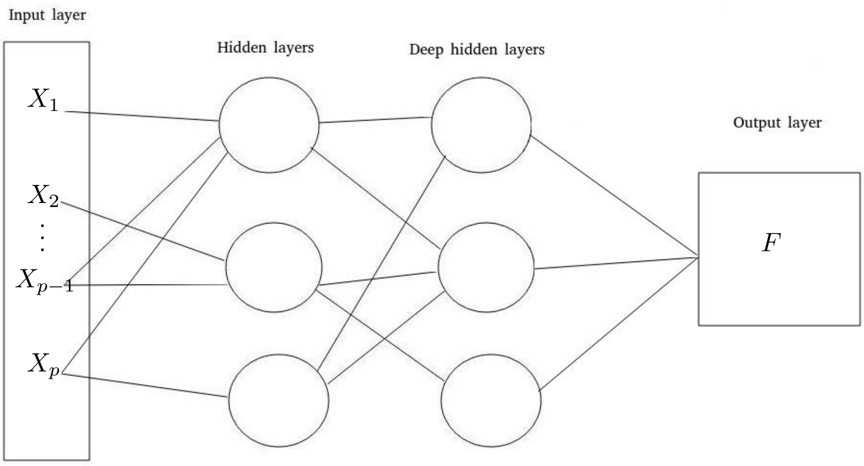


Figure 3: The procedure in the neural networks method explained. The input (the predictors $X_1, ..., X_p$) is transformed into one prediction F by using one or more hidden layers and one or more deep hidden layers. In each circle in this figure there is the multiplication with weights, adding of the bias and applying of the activation function happening, according to Equation (10).

gradient descent procedure by moving the error towards the minimum by following the gradient of the error. In each step one boost of following the gradient is applied. This leads then to updated weights and biases and we form new predictions $F_{m+1}(O_i)$ by using the weights and biases. By assuming that the weights can only be positive, we get a monotone neural network where the derivative of F with respect to the predictors is always positive.

The iterative procedure is applied for a given number of iterations to boost the predictions towards the observations. The predictions are not guaranteed to converge towards the observations, but we apply enough iterations to let the predictions become sufficiently close to the observations. In the probabilistic case, the error function is not differentiable at $F = O$, which can lead to difficulties in the gradient descent procedure. Therefore we change the error function in this case to the following smooth function:

$$\begin{aligned} E_F(O,F) &= \frac{1}{n}\sum_{i=1}^{n} \rho_q(h(F_i - O_i)), \\ h(x) &= \begin{cases} \frac{x^2}{2\epsilon} & \text{, if } |x| < \epsilon \\ |x| - \frac{\epsilon}{2} & \text{, if } |x| \geq \epsilon \end{cases} \end{aligned}$$

$h(x)$ is the Huber norm of x, which is a combination between an L^1 and L^2 norm. The Huber norm is applied with a small value for ϵ, so that the smooth version is almost equal to the non-smooth version. The smooth function has the advantage that it is differentiable everywhere. The value for ϵ decreases during the procedure, starting with 2^{-8} in the first step and decreasing with a factor 2 in each next step. Once the procedure has finished, we can use the final weights and biases to make predictions on a new set of predictors by plugging in the new predictors in the input layer and the values that come out in the output layer are the predictions.

To reduce the chance for overfitting we apply a penalized version of E_F, that penalizes for making the weights too large in the hidden layers:

$$E_F^{\text{pen}}(O,F) = \frac{1}{n}\sum_{i=1}^{n} \rho_q(h(F_i - O_i)) + \frac{\lambda}{W}\sum_{j=1}^{W} w_j^2, \tag{11}$$

where $\{w_j, j = 1, ..., W\}$ are the weights used in all the (deep) hidden layers. λ controls how strong the penalization is. Because λ doesn't depend on j, all the weights are equally penalized.

A second improvement that is applied to reduce the chance for overfitting, is applying the procedure in each step only on a subset of the data, just like in the gradient boosted regression trees method. The subset is randomly chosen by bootstrap samples and is therefore different each time. Consequently, in each step we only update the predictions for the observations that are in the subset that was chosen. By the tree-based methods the size of the subset is given and then the elements in the subset are randomly chosen, but for the neural networks method both the size and the elements in the subset are randomly chosen.

In the neural networks method we fit for different quantiles separately and that can cause crossings between quantiles. To solve this, we apply a composite version where we update the weights and bias for all quantiles $q_1, .., q_Q$ at the same time by using the error function:

$$\tilde{E}_F^{\text{pen}}(O, F) = \frac{1}{Q} \sum_{q=1}^{Q} E_F^{\text{pen}}(O, F)$$

Now for all the quantiles we update the weights and biases in the same way. Taken together with the monotonicity by requiring positive weights, we get the monotone composite extension of the neural networks method.

This extension of the neural networks method is applied in R using the package qrnn (Cannon (2018b)).

2.4 Scoring metrics for verification

To test whether our corrected forecast really improves the raw forecast, we use skill scores that express the relative improvement in a single number. First we have to define some scoring value A that compares the forecasts with the observations. This can for example be the sum over the errors between the forecasts and the observations. Then the skill score SS for one forecast with respect to a reference forecast is:

$$SS = \frac{A - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} \tag{12}$$

Here A is the score of the forecast we look at, A_{ref} the score of the reference forecast and A_{perf} would be the score of a perfect forecast. Using Equation (12) we see that higher skill scores are better with a maximum at 1 for perfect forecasts. To see if our corrected forecast improves the raw forecast we can use this skill score in two ways. We can compare them directly by taking the raw forecast as the reference forecast. Then a skill score higher then 0 indicates improvement. But we can also compare them indirectly by comparing the corrected forecast with a climatological forecast, which is a forecast that always assumes the climatological value, and comparing the raw forecast with the same climatological forecast. This yields then two skill scores, $SS_{\text{for\_corr}}$ and $SS_{\text{for\_raw}}$, and we have an improvement if the former is higher then the latter.

In this research we look at different scoring metrics to verify our forecasts, namely mean absolute errors, root mean squared errors, continuous ranked probability scores, Brier scores, reliability diagrams and potential economic values. These metrics are discussed in the following sections.

2.4.1 Mean absolute error and root mean squared error

For verifying deterministic forecasts the most generally used metrics are the mean absolute error (MAE) and the root mean squared error (RMSE), defined by:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |F_i - O_i|, \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (F_i - O_i)^2}, \tag{14}$$

where $\{O_i, i = 1, ..., n\}$ are the observations and $\{F_i, i = 1, ..., \}$ the forecasts corresponding to the observations. The MAE corresponds to the mean of the errors between the observations and the forecasts and the RMSE corresponds to the variance of the errors. We can calculate the MAE or RMSE skill score by calculating the MAE or RMSE for our forecast, a reference forecast and the perfect forecast and plugging these in in Equation (12). The perfect forecast is equal to the observations and has a MAE and RMSE of zero.

2.4.2 Continuous ranked probability score

The continuous ranked probability score (CRPS) is an extension of the root mean squared error for probabilistic forecasts. This metric is defined by:

$$\text{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} \left(G_{F_i}(y) - G_{O_i}(y)\right)^2 dy \tag{15}$$

Here G stands for the cumulative density function. For the observations this is an indicator function, $G_{O_i}(y) = \mathbb{1}_{y \geq O_i}$, and for the forecasts a smooth function going from 0 to 1. This is visualized in Figure 4. We can see the CRPS as an integrated version of the RMSE over the whole domain where the probabilistic forecast produces positive probabilities. We can calculate the continuous ranked probability skill score by plugging in the CRPS of our forecast, a reference forecast and the perfect forecast (for which the CRPS is zero) in Equation (12).

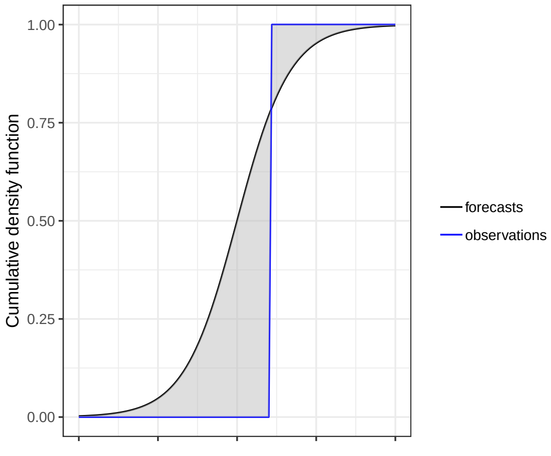


Figure 4: The CRPS (Eq. 15) is calculated by integrating the squared differences between the 2 graphs

2.4.3 Brier score

Another way to verify the probabilistic forecasts is to transform the continuous predictand first to a binary predictand using some threshold T. This means that every observation O_i is transformed to the indicator function $\tilde{O}_i = \mathbb{1}_{O_i \leq T}$. For the forecasts F_i we look at the probability mass that is below the threshold: $\tilde{F}_i = \mathbb{P}(F_i \leq T)$. This leads to single values for the transformed observations and forecasts. The transformed observations are 0 or 1 depending on if they are below the threshold and the transformed forecasts are a value between 0 and 1, standing for the probability of not exceeding the threshold. With these transformed observations and forecasts, we define the Brier score (BS) as follows:

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{O}_i - \tilde{F}_i)^2 \tag{16}$$

This score looks, just like the continuous ranked probability score, at the squared absolute differences between the observations and forecasts. Also for the Brier score, we calculate the Brier skill score by plugging in this definition in Equation (12). Then the perfect forecasts are the transformed observations with a Brier score of 0.

2.4.4 Reliability diagram

With the transformed observations and forecasts described in Section 2.4.3 using the threshold T, we can also take a look at the reliability diagrams. For this we measure how well the probabilities for the forecasts match with the observations. We split the probability range [0,1] into a number of bins B. Then for each bin $[\frac{b}{B}, \frac{b+1}{B}]$ we define the set of indices $I_b = \{i|\tilde{F}_i \in [\frac{b}{B}, \frac{b+1}{B}]\}$. The observed relative frequency ORF for each b is defined as the average of the observations in the set I_b:

$$ORF(b) = \frac{1}{|I_b|}\sum_{i=1}^{n} \tilde{O}_i \mathbb{1}_{i\in I_b}$$

This is also the fraction of observations in the set I_b that are equal to 1 and for a perfect probabilistic forecast this should be equal to the midpoint of the bin $[\frac{b}{B}, \frac{b+1}{B}]$ in the limit of an infinite number of observations. Therefore we compare for our forecast the midpoint of the bin with the observed relative frequency and by doing this for all the bins, we form the reliability plot with the midpoints of the bins on the x-axis and the observed relative frequencies on the y-axis. We connect the points with straight lines and the closer this resulting line is to the diagonal 1:1 line, the better our forecast is. This means that the reliability plots are also a way to verify the forecasts, but they only look at the forecasts and not at a climatological forecast or a perfect forecast.

2.4.5 Potential economic value

The potential economic value is, as its name says, a value that describes the potential economic impact of the forecasts. It is defined by a simple economic model concerning costs and losses and is described in Richardson (2000). It starts by looking at a certain event. This can be exceeding a certain amount of rain or exceeding a certain wind speed. For this research we look at the event of not exceeding a certain amount T of global radiation. Then when the event is forecast to happen (less than T amount of radiation forecast), users of solar energy can decide to apply some costs C. This is for example buying alternative sources of energy beforehand. Then if the event actually occurs or not, does not matter anymore, because the costs are already applied. When the event actually occurs we call it a 'hit', and otherwise a 'false alarm'. On the other hand, if the event is not forecast to happen, users expect to produce enough energy and do not apply the costs. If the event then also does not happen, the right choice has been made of not applying the costs and there is no loss suffered (this is called a 'correct rejection'). If the event happens, then some loss L is suffered by, for example, having to buy alternative sources of energy last minute (which will be more expensive than buying it beforehand). Also the loss L could represent paying some fine, because of not supplying enough energy to the electricity grid. This last scenario is called a 'miss'. The four scenario's are illustrated in Table 1:

	Observed	Not observed
Forecast	Hit → C	False alarm → C
Not forecast	Miss → L	Correct rejection → 0

Table 1: The 4 different scenario's in the economic model corresponding to the event that the amount of radiation does not exceed a certain value T.

For this model the assumption is made that $C < L$, otherwise it is always better to not pay the costs C. The expected expense is C if the event is forecast and $L \cdot p_O$ if the event is not forecast, where p_O is the probability of observing the event given that the event is not forecast. It is therefore better to apply the costs when $C < L \cdot p_O$ or $\frac{C}{L} < p_O$.

Suppose now that we not have a yes/no forecast for the event, but a probability p_F of forecasting the event. This probability is calculated from the full probabilistic forecast by applying the threshold T on it according to the procedure described in Section 2.4.3. We make the assumption that the forecast probability is reliable, meaning that the forecast is calibrated. Then it holds that $p_O = p_F$ and we apply the costs when $\frac{C}{L} < p_F$. For users with a fixed ratio $\tilde{T}$ between the costs and losses, we can see the forecast as a 'yes' forecast (the event is forecast to occur) when $\tilde{T} < p_F$ and else as a 'no' forecast. The expected expense is C when the yes/no forecast in combination with the observation leads to a hit or a false alarm. The expected expense is L with a miss and 0 with a correct rejection.

When we have a set of forecast probabilities $\{\tilde{F}_i, i = 1, ..., n\}$ corresponding to observations $\{\tilde{O}_i, i = 1, ..., n\}$ (similar to the notation in Section 2.4.3, meaning that the transformation with threshold T has been applied), the expected expense E_{for} for the set of forecasts is:

$$E_{\text{for}} = (H + FA)C + ML, \tag{17}$$

where H, FA and M are the frequency of forecasts that lead to respectively a hit, false alarm and miss. For a climatological reference forecast the expected expense can also be calculated. The climatological reference forecast always gives the observed relative frequency ORF as forecast, defined by:

$$ORF = \frac{1}{n}\sum_{i=1}^{n} \tilde{O}_i$$

The expected expense E_{ref} for the climatological reference forecasts is:

$$E_{\text{ref}} = \begin{cases} C & \text{if } \frac{C}{L} < ORF \\ L \cdot ORF & \text{otherwise} \end{cases} \tag{18}$$

For perfect forecasts, equal to the observations, the loss never occurs and the expected expense E_{perf} is:

$$E_{\text{perf}} = C \cdot ORF \tag{19}$$

Using the expected expenses from our forecast, the climatological reference forecast and the perfect forecast, we calculate the potential economic value (PEV) in the same way as in Equation (12):

$$\text{PEV} = \frac{E_{\text{for}} - E_{\text{ref}}}{E_{\text{perf}} - E_{\text{ref}}} \tag{20}$$

By filling in the expressions for $E_{\text{for}}, E_{\text{ref}}$ and E_{perf}, we get:

$$\text{PEV} = \begin{cases} \frac{\frac{C}{L}(H+FA-1)+M}{\frac{C}{L}(ORF-1)} & \text{if } \frac{C}{L} < ORF \\ \frac{\frac{C}{L}(H+FA)+M-ORF}{(\frac{C}{L}-1)ORF} & \text{otherwise} \end{cases} \tag{21}$$

The potential economic value is a function of the cost-loss ratio and the frequencies of hits, false alarms and misses of our forecasts. The value is 1 when our forecasts are perfect and 0 when our forecasts have equal performance as the climatological reference forecast. In the results section, we calculate the potential economic value for a range of values for the cost-loss ratio between 0 (no costs) and 1 (costs equal to the losses). Each cost-loss ratio corresponds to a different value for $\tilde{T}$ (equal to the cost-loss ratio), used in calculating the expected expense for our forecasts. Therefore we calculate the potential economic value not only for a set of cost-loss ratios, but at the same time also for the same set of values for $\tilde{T}$. Then the potential economic value for each cost-loss ratio is the maximum over the potential economic values for the different values of $\tilde{T}$. This gives us a set of potential economic values for the different cost-loss ratios, where each cost-loss ratio corresponds to a certain part of the users of solar energy.

3 Data and fitting procedure

In this section we describe the data that we use and the fitting procedure that is applied to achieve the results, which are presented in the next section. The goal of the research is to improve the forecasts of the solar radiation that reaches the Earth. We investigate the global (or total horizontal) solar irradiance (in W/m^2), which consists of a direct component and a diffuse component. The direct radiation is the radiation that directly comes from the sun on a plane perpendicular to the solar radiation beams. The diffuse radiation is the radiation from the sun that is scattered by for example clouds or aerosols, but still manages to reach the Earth's surface. The global radiation G can be described in terms of the direct radiation D and the diffuse radiation B by the following expression:

$$G = D \cdot \cos z + B,$$

where z is the solar zenith angle, which is the angle between the plane perpendicular to the Earth's surface and the position of the sun in the sky.
The forecasts of the global radiation that we use are deterministic and we improve them both in a deterministic way and in a probabilistic way. The probabilistic approach gives more information about the forecasts by not just giving a single number of the most probable value of global radiation, but by giving a list of quantiles from the underlying distribution of the global radiation. The quantiles give a number of distinct values for the forecast of global radiation, each with an associated fixed probability level. The quantiles also give information about the uncertainty in the forecast. If the quantiles in the list are close to each other, we can be quite certain about the forecast. However when the quantiles

are not so close to each other, we know that there is a lot of uncertainty in the forecast. We can also get information about the skewness of the forecast by looking at the median. If this is approximately in the middle between the lowest and largest quantile, then the distribution of the forecast is close to symmetrical. In contrast, if the median is close to either the lowest or the largest quantile, then the distribution is asymmetrical. Our research mostly focuses on the probabilistic forecasts, because they give much more information about the forecast global radiation. We still show some deterministic verification scores, based on the median of the probabilistic forecast. The probabilistic forecasts are formed by 49 quantiles: $\{0.02, 0.04, ..., 0.5, ..., 0.96, 0.98\}$. The 25th quantile is then the median and the other quantiles are evenly spread out around the median.

3.1 Model data and observations

To produce the results shown in the next section, we apply the methods that were described in the previous section to the data set we have. This data set consists of the predictand (the global radiation observations) and the potential predictors, with which the methods produce new predictions for the predictand. The period for which we use data consists of April 1, 2016 through March 16, 2018.

3.1.1 Global radiation observations

The predictand is in our research the hourly observations of the global radiation measured at 30 weather stations in the Netherlands by the Netherlands network of automatic weather stations. These observations are retrieved from the KNMI website (KNMI (2018)). The locations of the stations are shown in Figure 5. The observation for each hour is defined and measured as the average global radiation over the previous hour.

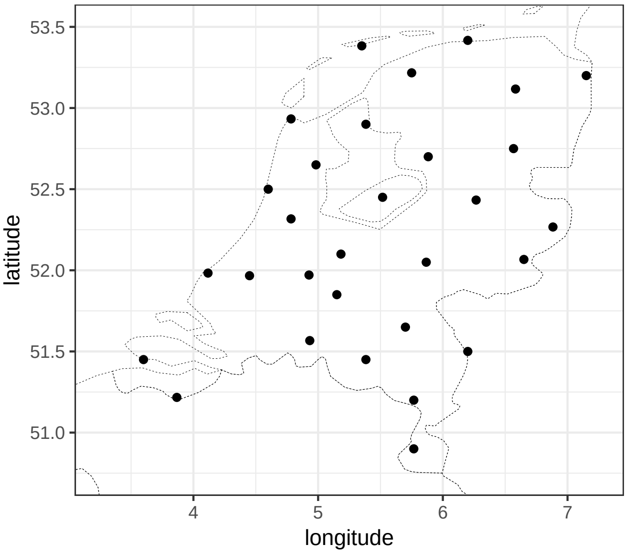


Figure 5: The locations of the stations

3.1.2 Potential predictors

The set of potential predictors we use consists of atmospheric temperature, humidity and different cloud, radiation and aerosol variables. In Table 3 there is an overview of the potential predictors. Most of them (marked with *) come from the output of the NWP model HARMONIE 38, which is version 38 of the high-resolution weather model at KNMI. The HARMONIE model produces forecasts of atmospheric variables by applying physical equations about the (thermo)dynamics in the atmosphere starting from the initial conditions of the variables. The dynamical equations are discretized over a grid and for effects taking place inside grid boxes of the model (like cloud formation) there are physical parametrizations describing these local effects. The initial conditions for the model are a combination of observations and the model output from the previous run valid at the current time. The optimal combination between the observations and model output for the initial conditions is found by a data assimilation procedure called

3DVAR (ECMWF (2002)). The boundary conditions in space for the 3DVAR procedure are provided by the model output from the ECMWF model, which produces forecasts for the whole global domain. More detailed information about the HARMONIE model can be found in Bengtsson et al. (2017). The aerosol predictors (marked with ** in Table 3) come from the CAMS forecast. This is the Copernicus Atmosphere Monitoring Service, part of ECMWF, that delivers forecasts for aerosols and other particles (for example ozone) in the atmosphere over the entire Earth. The reason that we use this additional source of aerosol forecasts, is because the HARMONIE model uses climatological averages for the aerosols and other particles. By using the CAMS forecasts, we not only have the climatological averages from HARMONIE, but also values for the aerosols that change over time and space. We expect this to lead to better results for the solar radiation forecasts, because the aerosols influence the amount of radiation from the sun that reaches the Earth, by reflecting part of it back into space. In Table 2 we have listed some facts about the HARMONIE and CAMS models.

	Harmonie 38	CAMS
Type of forecast	Deterministic	
Starting time	0 UTC	
Time domain	0-48 hours	0-120 hours
Spatial domain	[0°, 11.063°]x[49°, 55.877°]	Global
Time resolution	hourly	3-hourly
Spatial resolution	2.5km x 2.5km	0.5° x 0.5°

Table 2: Information about the HARMONIE and CAMS models.

Some dates in the period we look at have no model output from HARMONIE and those dates are removed from the period. For CAMS we only use the forecasts that fall in the time and space domain of HARMONIE. For the time and space resolution we have made them the same as for HARMONIE by taking for each point in the space and time resolution from HARMONIE, the value from the CAMS forecast that is the closest to the point both in space and time. The CAMS forecasts are in a practical setting not fast enough available to be used in the fitting procedure and therefore we use for each date the CAMS forecast from the previous date at 0 UTC and then look at time span of 24-72 hours instead of 0-48 hours.

(a) The potential predictors with * are from HARMONIE and the potential predictors with ** are from CAMS. SURF stands for the Earth's surface. The layers indicate the lower (0-2000m), middle (2000-6000m), upper (6000m-TOA) and total (0-TOA) part of the atmosphere, where TOA is defined in the text below.

	Potential predictors
Temperature*	T_{layers}
Humidity*	RH_{layers}
Radiation*	Global, Direct (surf/TOA), NCS_Global (surf/TOA)
Clouds*	Rain, CC_{layers}, CW_{layers}, PW_{layers}
Particles**	AOD_{500}, Ang_exp, Ozone
Time/place	Lat, Lon, DoY, CosZen, DistToCoast, DistToWater, DistToInland

(b) The abbreviations of the potential predictors explained. Netto clear sky stands for the incoming minus the outgoing radiation under a clear sky.

T	Temperature
RH	Relative humidity
NCS	Netto clear sky radiation
CC	Cloud cover
CW	Cloud water
PW	Precipitable water/water vapor
AOD_{500}	Aerosol optical depth at 500 nm
Ang_exp	Angstrom exponent
Lat	Latitude
Lon	Longitude
DoY	Day of the year
CosZen	Cosine of zenith angle

Table 3: The potential predictors

For each of the potential predictors with 'layers' as a subscript, we have them for the lower part (0-2 km), middle part (2-6 km) and upper part of the atmosphere (6 km and higher) and also an aggregated version of the total atmosphere (except for temperature and relative humidity, where the total atmosphere component is missing). The HARMONIE output for the cloud water consists of 65 levels in the atmosphere and we applied a weighted average to reduce these to the four layers (lower, middle, upper and total atmosphere):

$$CW_j = \sum_{\substack{i=1,\\ h(i)\in L_j}}^{65} \frac{h(i)-h(i-1)}{h_j(\max)-h_j(\min)} CW(i), j = 1, ..., 4, \tag{22}$$

where $CW(i)$ and $h(i)$ are respectively the cloud water (in kg/m^2) and the height (in m) of level i and L_j (j = 1,..,4) denotes the range of heights (in m) of layer j, defined by $L_1 = [0, 2000]$, $L_2 = (2000, 6000]$, $L_3 = (6000, \mathrm{TOA})$ and $L_4 = [0, \mathrm{TOA})$, where TOA stands for the top of the atmosphere, defined as the height with an air pressure of 0.01 hPa. The quantities $h_j(\max)$ and $h_j(\min)$ are defined by:

$$h_j(\max) = \max_{i, h(i)\in L_j} h(i),$$
$$h_j(\min) = \min_{i, h(i)\in L_j} h(i)$$

For the temperature and humidity the HARMONIE output consists of 11 pressure levels and we reduce these to the three layers (lower, middle and upper atmosphere) by applying a weighted average in the same way as in Equation (22).
The precipitable water is not directly output from HARMONIE, but by using the temperature and humidity of the 11 pressure levels together with the formulas in McRae (1980), we calculate the precipitable water at 11 pressure levels and then apply a weighted average as in Equation (22) to reduce it to the four layers (lower, middle, upper and total atmosphere).
For all radiation predictors we have that the values represent the average radiation over the previous hour. For the rain predictor the values represent the total amount of precipitation that has fallen over the previous hour. For all the other potential predictors the values are valuable at the exact moment in time.
The time/place predictors are not output from HARMONIE or CAMS, but are calculated and used as potential predictors. These consist of the latitude and longitude of the station locations, the day of the year, the distance to the coast (minimum distance to one of the seas), distance to water (minimum distance to the sea or to one of the lakes) and distance to inland (the distance to the intersection point of the borders of the Netherlands, Belgium and Germany). The cosine of the zenith angle is also used as a potential predictor and is calculated using Michalsky's algorithm (Michalsky (1988)).

3.2 Fitting procedure

We apply all the methods described in sections 2.2 and 2.3 to the data set we have. Specifically, we apply three methods for the parametric approach. We fit two distributions for the predictand, the gamma distribution (denoted by GA) and the truncated normal distribution (denoted by NOtr). The truncated normal distribution is a modified version of the normal distribution defined only on $\mathbb{R}^+$. Therefore both the gamma and truncated normal distribution only have probability mass on the positive part of the real line, which suits for predicting the (always positive) radiation. The probability density functions of the gamma and truncated normal distribution are defined as follows:

$$p_{GA}(x) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{x^{\frac{1}{\sigma^2}-1} e^{-x/(\sigma^2\mu)}}{\Gamma(\frac{1}{\sigma^2})}$$
$$p_{NOtr}(x) = \frac{p_{NO}(x)}{CDF_{NO}(\infty) - CDF_{NO}(0)},$$

Where the probability density function p_{NO} and cumulative density function CDF_{NO} of the normal distribution are defined by:

$$p_{NO}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$
$$CDF_{NO}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{y^2}{2}} dy$$

Both the gamma and the truncated normal distribution are defined by two parameters: μ and σ. These two parameters are fitted by using Equation (5). The third parametric method is quantile regression, which we denote by QR. For this method we add some random noise, generated from a normal distribution with mean zero and a variance of 0.001, to the predictors to prevent the matrix with the predictors from becoming singular. This has to be done, because in the minimization of the WSR (Eq. 6) the matrix with the predictors gets inverted. A singular matrix leads to an error in the algorithm, but with the added (small) random noise this problem does not occur anymore.

We also apply the four non-parametric methods, described in Section 2.3. These are denoted by QRF for the quantile regression forests, GRF for the generalized random forests, GBRT for the gradient boosted regression trees and ANN for the artificial neural network. We compare the 7 methods mutually and with the raw forecast.

The fitting of the methods is done per lead time, because systematic errors increase with lead time, and for all stations and dates at once. For the stations we take the predictors at the gridpoints closest to them. Then we can compare predictions made by the methods with the observations at the stations and apply verification measures to the predictions. Once the methods have been fit, we can compute predictions for each gridpoint instead of only at the gridpoints closest to the stations.

Before fitting the methods we divide all the radiation predictors and the predictand by the clear sky radiation, the global radiation that would have been measured if the sky would have been cloud free. We do this to filter out the large seasonal and diurnal cycle of radiation. The radiation divided by the clear sky radiation is called the clear sky index. To calculate the clear sky radiation we use the European Solar Radiation Atlas (ESRA) clear sky model. This model is explained in detail by Rigollier et al. (2000). It uses the zenith angle as an input, which we calculate using Michalsky's algorithm. Another input is the Linke turbidity factor, for which we use monthly average values for De Bilt, see Remund et al. (2003). Further inputs are the atmospheric pressure, which we take equal to the mean atmospheric pressure at sea level of 1013 hPa, the total solar irradiance constant of 1367 W/m^2 and the distance Earth-sun of 1 AU. Because the radiation predictors and the predictand are average values over an hour, we do the same for the clear sky radiation. This is done by calculating the zenith angle 60 times, at the 60 minutes in the hour, and then averaging the zenith angles to obtain the average zenith angle during that hour. This forms the zenith angle input and together with the other inputs we calculate the hourly average clear sky radiation with the algorithm described in Rigollier et al. (2000).

In the fitting procedure, for each lead time t we split the data set D_t into a training set S_t and a testing set $D_{t \setminus S_t}$. On the training set we train our methods to find relationships between the predictors and predictand. Then on the testing set we generate predictions using the relationships that were found in the training set and we compare these predictions with the actual observations to verify how close the predictions are to the observations. We change this training and testing set in such a way that every data point in the set D_t appears once in the testing set and twice in the training set. This is a 3-fold cross-validation procedure. We split the data set into the training and testing set based on groups of consecutive dates. All stations for one date appear either in the training or in the testing set.

In the fitting procedure we can either take the whole data set D_t and split it into training and testing sets or we can split D_t into the four seasons and then fit per season by making training and testing sets. The four seasons are the meteorological seasons, defined by the Winter {December, January, February}, the Spring {March, April, May}, the Summer {June, July, August} and the Autumn {September, October, November}. We investigate which approach leads to the best predictions. We only take dates and stations into account during daylight. This is done by removing all combinations of dates and stations that have a clear sky radiation lower than some threshold. For our research we take this threshold equal to 20 W/m^2, so if date/station combination has a clear sky radiation lower than 20 W/m^2, we consider it to be in the night and remove it. Only for the lead times +5h till +19h and +29h till +43h we keep enough dates and stations (defined by at least 50), when the night values are removed to continue the fitting procedure. For the predictors we investigate the influence of averaging them over multiple lead times and/or grid points, to see which approach leads to the best predictions.

To calculate the (skill) scores we combine the predictions over all testing sets. Then we calculate the score over a chosen subset of the predictions using the procedures described in Section 2.4. The chosen subset can just be the whole set or for example one of the seasons. The reference forecast in Equation (12) is also calculated on the chosen subset. Here we take the reference forecast to be the sample climatology. For a deterministic reference forecast this is the mean of the observations in the subset. For a probabilistic forecast it consists of quantiles drawn from the observations in the subset. These are the same 49 quantiles as we use in forming our probabilistic forecasts. For the Brier score the sample climatology is the mean of the transformed observations in the subset.

In the results section we show both a few examples of the predictions that are made as well as verification scores and skill scores that are calculated for the whole subset. The predictions are multiplied by the clear sky radiation to get the actual radiation, which causes the predictions and scores to be in W/m^2 and become better interpretable. For the skill scores we do not multiply with the clear sky radiation. Then the skill scores are computed directly after the fitting procedure instead of first applying a transformation on the predictions.

For achieving the best results for each method we find the optimal settings in each method and also in the fitting procedure, which consists of averaging the HARMONIE predictors over space and time and fitting per season or for the whole data set at once. The procedure for finding the optimal settings is found in Appendix A. To summarize the results in Appendix A, we find that averaging the HARMONIE predictors in space over a 9x9 block and in time over 3 lead times is the best approach. Also fitting per season separately produces better results than fitting the complete data set at once. The optimal settings in each method are listed in Table 4:

	GA	NOtr	QR	ANN		QRF	GRF	GBRT
Steps for μ or quantiles	5	5	5	-	Trees	500	500	100
Steps for σ	1	1	-	-	Minimal nodesize	5	5	5
Hidden layers	-	-	-	1	Sampling fraction	1/2	1/2	1/2
Deep hidden layers	-	-	-	3	Predictor fraction	1/3	1/3	-
Iterations	-	-	-	10	Depth of trees	-	-	1
Penalty	-	-	-	0	Learning rate	-	-	0.1

Table 4: The optimal settings in each method. The description for the parameters of the methods is found in Appendix A.

4 Results

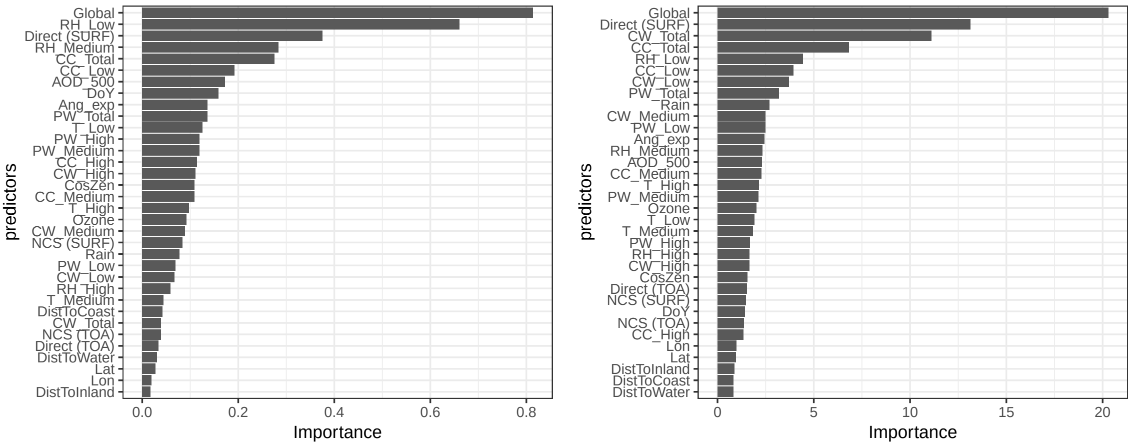
In this section we use the optimal settings in both the fitting procedure and in the hyperparameters of the post-processing methods to produce the predictions of global radiation. First we show the importance of the potential predictors. Then we show the verification of the median of the forecasts (MAE and RMSE), some case studies for showing the predictions on selected days, and scoring metrics calculated for all predictions. The selected days clearly show the added information of the probabilistic forecasts in comparison to deterministic forecasts. The scoring metrics consist of the continuous ranked probability skill score, the Brier skill score and reliability diagrams for comparing the methods. Furthermore, the potential economic value is shown for a more objective comparison between the raw forecast and the methods.

4.1 Importance of predictors

In producing the probabilistic forecasts, the 7 methods use information from the potential predictors. Some predictors are more important or used more than other predictors. To see how important the different predictors are, we show the importance measure for one parametric (GA) and one non-parametric (QRF) method in Figure 6. For the GA method the importance measure of some predictor in one fit is defined as 1 if the predictor is chosen in the fit (either in μ or σ) and 0 if not. Consequently, we consider the predictor to be important if it is in the fit. For the QRF method, the importance measure is defined differently. For some predictor X_j it is in one tree defined as the weighted sum over the homogeneities defined by Equation (8) for the splits where X_j is used to make the split. The weights are $\frac{|S|}{n}$, with $|S|$ the size of the set to make the split on (the same S as in Equation (8)) and n is the total number of observations. The importance measure for one fit (i.e. one forest of trees) is then the average over the importance measure values from the trees in the forest.

For both the GA and QRF method the importance measure values shown in Figure 6 are averages over all fits. These consist of all lead times, the four seasons and the three cross-validation sets.

The ranking of the predictors is more informative than the actual importance measure values they have. For the ranking, we see that the HARMONIE forecast of global radiation, called the raw forecast, is the most important predictor for both methods. This is what one would expect, because the raw forecast already accounts for all kinds of effects from other meteorological variables such as clouds. The



(a) GA method

(b) QRF method

Figure 6: Importance of the potential predictors in the GA (a) and QRF (b) methods, averaged over all fits.

raw forecast does not account optimally for these effects and therefore other predictors are also quite important, such as the total amount of cloud water, the total cloud cover and the relative humidity. The rest of the predictors are less important, but none of them have zero importance, so they are all used in producing the predictions. The QRF method has the advantage that it uses all predictors in all fits, whereas the GA method has a limit on the amount of predictors it uses to prevent overfitting. Therefore the GA method also uses all predictors, but not in all fits. The advantage that the QRF method has over the GA method holds more generally in that the non-parametric methods have the advantage over the parametric methods of using all predictors in all fits.
Next, it is good to note that for almost all predictors for which we have a low, medium and high component, the low component is more important than the other two. This means that the lower part of the atmosphere (0-2 km) has more influence on the global radiation that reaches the Earth's surface than higher parts of the atmosphere. For the aerosol predictors from CAMS we see that they are of medium importance for QRF and even more important for GA (especially for the aerosol optical depth and Angstrom exponent). In general we can say that the cloud related predictors are more important than the aerosol predictors, but there are certainly also cases in which the aerosol predictors turn out to be important. We will see an example case for this statement in Section 4.3.1.

4.2 Deterministic verification

For verification of the probabilistic forecasts 49 quantiles are used, but we first look deterministically at the forecasts by considering the median (the 25th quantile). We compare the medians with the observations by computing the MAE (Eq. 13) and the RMSE (Eq. 14) for the different methods and list them in Table 5. To make the numbers more interpretable, the MAE and RMSE are calculated over the actual radiation and not over the clear sky indices. This is done by multiplying all the forecasts with the clear sky radiation. In Table 5 the MAE and RMSE of the sample climatology and the raw forecast are also listed to see the improvement of the median forecast in both the MAE and RMSE. We also compare the methods mutually. The different rows in the table represent the different subsets (the seasons) on which the MAE and RMSE are calculated. All the values are averages over the stations and we show 2 lead times (+12h and +36h).

We can see in Table 5 that the MAE of the raw forecast are already much lower than the sample climatology and this MAE is (slightly) further reduced by all post-processing methods. The MAE in spring and summer is higher than in winter and autumn, because there is more radiation in spring and summer and also more convection, that can cause cloud formation, which makes the global radiation harder to predict. In spring and summer the methods reduce the MAE of the raw forecast more than in winter and autumn. The methods are able to account at least partly for the effects of convection. Comparing lead time +12h and +36h we see that the MAE increases with lead time, as expected, because systematic errors increase with lead time. We also see that the MAE of the raw forecast increases more

(a) The MAE at lead time +12h.

	CLIM	RAW	GA	NOtr	QR	QRF	GRF	GBRT	ANN
Winter	99.24	52.59	46.68	47.52	45.23	46.59	**45.16**	45.89	46.95
Spring	194.25	117.32	104.99	102.97	100.80	100.66	**99.93**	101.18	105.69
Summer	194.61	142.10	116.65	115.58	**112.43**	114.07	113.79	114.09	117.46
Autumn	152.25	82.67	67.21	67.34	**66.05**	66.35	66.21	66.45	68.41

(b) The MAE at lead time +36h.

	CLIM	RAW	GA	NOtr	QR	QRF	GRF	GBRT	ANN
Winter	101.59	60.12	56.53	54.02	**51.59**	55.19	53.05	52.99	53.79
Spring	196.57	127.52	114.36	110.24	109.05	109.33	**108.45**	109.39	112.68
Summer	193.90	154.21	127.12	126.10	123.67	123.88	**123.51**	124.15	129.38
Autumn	151.31	90.48	76.68	74.78	**72.87**	74.58	74.25	75.34	77.43

(c) The RMSE at lead time +12h.

	CLIM	RAW	GA	NOtr	QR	QRF	GRF	GBRT	ANN
Winter	121.56	72.13	63.43	63.37	**62.72**	65.05	63.73	63.62	66.22
Spring	227.90	163.81	138.71	137.50	138.34	136.53	**136.40**	137.37	145.57
Summer	226.14	194.69	151.35	**151.05**	151.65	151.57	151.34	151.79	157.65
Autumn	180.84	117.30	91.89	91.14	91.21	91.27	91.37	**90.95**	95.00

(d) The RMSE at lead time +36h.

	CLIM	RAW	GA	NOtr	QR	QRF	GRF	GBRT	ANN
Winter	124.09	83.17	74.56	71.78	**71.66**	76.86	74.44	72.26	75.67
Spring	230.39	175.83	148.12	**145.59**	148.38	146.64	147.17	146.50	154.22
Summer	225.32	209.99	163.79	162.24	164.16	**162.12**	162.51	163.54	170.47
Autumn	179.86	127.05	100.96	**98.46**	98.47	99.19	99.56	99.99	104.91

Table 5: The MAE (Eq. 13) and RMSE (Eq. 14) for the different methods calculated over the four seasons at lead times +12h and +36h (averaged over all stations). The lowest (best) scores are highlighted. CLIM stands for the sample climatology, RAW for the raw forecast and the rest of the abbreviations are explained in Section 3.2.

than that of the other methods. The loss in accuracy of the raw forecast with increasing lead time is reduced by the methods. The best performing method is highlighted in each row and we see that the QR and GRF method perform best. It depends both on the season and the lead time which one of them performs slightly better.

The RMSE results are similar. The raw forecast again greatly improves on the RMSE with respect to the sample climatology. The methods reduce this RMSE even more. We see some seasonal dependence. In spring and summer the RMSE is more reduced by the methods than in winter and autumn. This means that the methods account better for the effects of convection. Also the increase in RMSE from lead time +12h till +36h is lower for the methods than for the raw forecast. Again the loss of accuracy of the raw forecast with increasing lead time is reduced by the methods. For the MAE the QR and GRF methods perform best, but for the RMSE there are more methods that perform best for a certain season or lead time. Only the GA and ANN method are in none of the cases best performing. The rest of the methods perform best in at least one of the cases.

4.3 Probabilistic verification

For the deterministic scoring metrics it was already clear that the post-processing methods improve the raw forecast, but by mutually comparing the methods it appears that there is no clear best performing method. Now we present the verification of the probabilistic forecasts. First we show for a few selected cases the added information of estimating the uncertainty in the forecast. Then we present results from a number of scoring metrics to compare the methods objectively. We finish with the potential economic value, for which we compare the methods with the raw forecast.

4.3.1 Predictions on selected cases

Before calculating scoring metrics for the predictions, we look at a number of forecast examples for station Cabauw. For this we have selected a number of interesting cases, each case interesting in its own way. We start by studying the influence of clouds on the radiation. A good reference point is a clear sky day, which is a day without clouds in the atmosphere. We take the specific case of April 9, 2017, which was a clear sky day. We look at the forecast initialized at 0 UTC on the 9th of April and compare the predictions for the lead times +6h till +18h with the observations for 6 UTC till 18 UTC. We have done this for two methods: the parametric GA method and the non-parametric QRF method. In Figure 7 we have plot the observations, the raw forecast, the medians of the GA and QRF method and the range of uncertainty. The gray area denotes the probability distribution from the 0.02 quantile till the 0.98 quantile. We see that the raw forecast is already very accurate and this can be seen on all clear sky days. Although the raw forecast is very good in predicting clear sky days, there is still a small negative bias (observations higher than raw forecast). Both GA and QRF are not able to reduce the bias and have instead a small positive bias for the median. Looking at the uncertainty in the forecasts, we see that both methods are very certain (the QRF method even more than the GA method) about the forecast radiation, because the gray bands are small. The bandwidth is around 200 W/m^2 for the GA method and 100 W/m^2 for the QRF method in the middle of the day. Both the raw forecast and the other two methods predict a high amount of radiation, which indicates a clear sky day. The GA and QRF method provide the additional information that it is (very) certain that a high amount of radiation is reaching the surface. Information about the uncertainty in the forecast is always very valuable for users, also when the forecast is very certain.

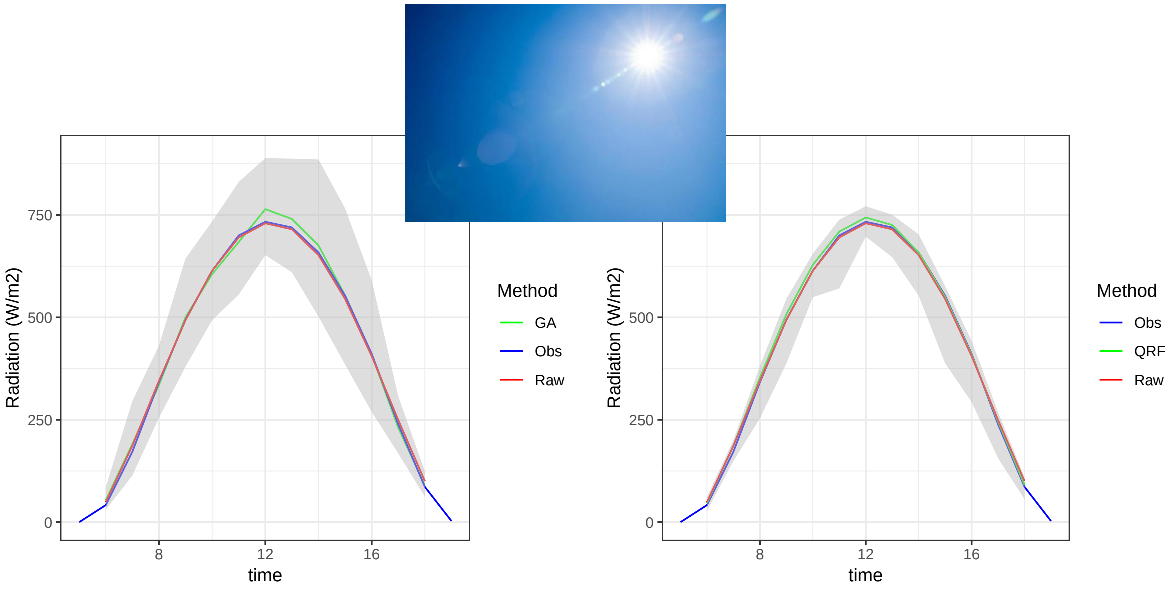


Figure 7: The case of April 9, 2017 at station Cabauw. Obs stands for the observations, Raw for the raw forecast, GA for the gamma distribution and QRF for the quantile regression forests method. The gray area denotes the probability distribution from the 0.02 until the 0.98 quantile.

The opposite of a clear sky day is a fully overcast day. For this we take the case of November 20, 2017. This day had clouds in the sky for the whole period between 5 UTC and 19 UTC, which is visible in the low amounts of observed radiation in Figure 8. We compare the GA and QRF method with the raw forecast for lead times +8h till +15h. We see that the raw forecast clearly underestimates the amount of radiation. The raw forecast has values around 10 W/m^2, which is unrealistically low, even on a fully overcast day. Both GA and QRF compensate for this bias in the median and also indicate that it is very certain that it will be a fully overcast day, with the highest quantile 0.98 of the distribution staying below 150 W/m^2. Only in the optical thickness of the clouds there is still uncertainty. The distributions of both the GA and QRF method are asymmetrical and are skewed towards higher (than the median) amounts of radiation than lower amounts. For this particular case this is realistic, because the median predicts a low amount of radiation (around 50 W/m^2) and consequently the distribution is skewed towards higher amounts of radiation. Both methods predict that this day is a fully overcast day with relatively high certainty and both medians are considerably better than the raw forecast.

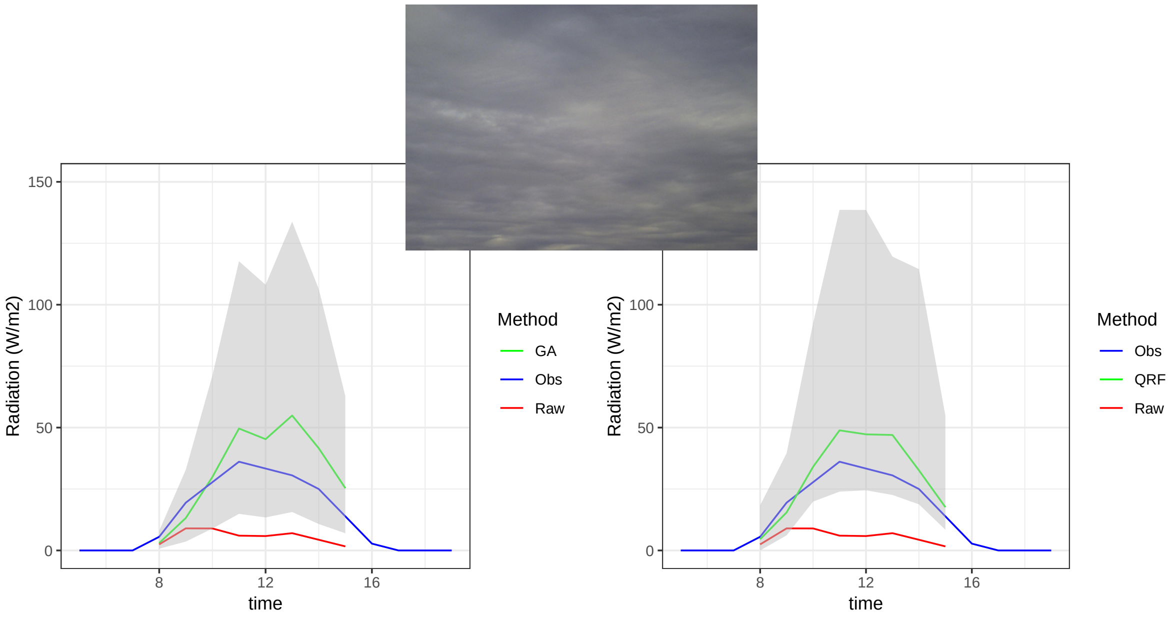


Figure 8: As Figure 7 but for November 20, 2017.

Days that are more difficult to predict include days that are not fully overcast and also not completely sunny, which we call a partly overcast day. As an example we have the case of April 23, 2017 in Figure 9. We see that the observed radiation is highly variable during the day. Comparing the observations with the forecasts for lead times +6h till +19h, we see that the raw forecast has a large bias for most lead times. Both the median of the GA and QRF method have reduced this bias, especially in the middle of the day. The middle of the day is also the part with the highest observed radiation (and consequently the highest power production in solar panels), so it is very important to have good predictions for this part of the day. In this case there is a lot of uncertainty in the forecast. The lowest quantiles predict a fully overcast day, while the highest quantiles predict a completely sunny day. This indicates that this is a day for which it is difficult to forecast the radiation that is going to be observed and therefore it is good to not fully trust the median of the distribution, but also be prepared for lower or higher amounts of radiation. In the Netherlands, partly overcast days occur regularly and so it is very helpful to look at the probabilistic forecast to see that there is a large uncertainty in the forecast of global radiation.

We also investigate the effect that aerosols have on the radiation forecast. We take the interesting case of October 17, 2017, which was a day without clouds. Although there were no clouds, there were many particles in the sky consisting of Sahara sand and dust and ash particles from forest fires. This caused the number of aerosols in the sky to become very large and those aerosols blocked a lot of radiation as can be seen in the observations in Figure 10. The raw forecast is based on climatological values for the aerosols and was not able to account for this large amount of aerosols. Therefore it predicted a clear sky day with a large amount of radiation, as can be seen from the large difference between the raw forecast and the observations. Both the median of GA and QRF reduce this bias slightly by predicting lower amounts of radiation. The medians are still close to the raw forecast, because the raw forecast (the global radiation) is the most important predictor as we have seen in Figure 6. The more interesting part of this case is to look at the distributions. For the GA method the distribution is close to symmetrical, which is not very informative. However the distribution of the QRF method is very asymmetrical towards lower amounts of radiation. This indicates that there is a lot of uncertainty in the forecast skewed towards lower amounts of radiation. The observations show that these lower amounts of radiation were also measured on this day and the observed radiation is in the middle of the day even below the 0.02 quantile. This day is extreme in the sense of a very high amount of aerosols and is a very challenging day to forecast global radiation due to that there is not a similar day in the training set. Comparing the GA and QRF method, we conclude that on this day the QRF method performs better, because an asymmetrical distribution is more realistic for this day.

We also show forecasts for the GA and QRF method where we have not used the aerosol predictors from CAMS in the fitting procedures. In general the aerosol predictors are less important than the cloud related predictors. But on this day where there are no clouds we can see the added value of the aerosol

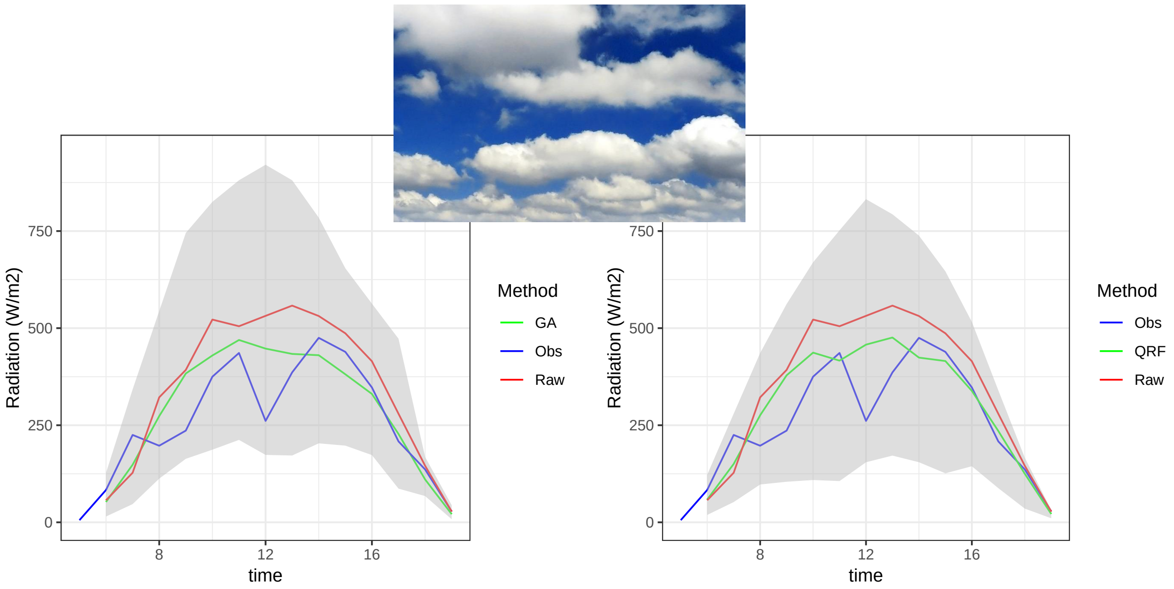


Figure 9: As figure 7 but for April 23, 2017.

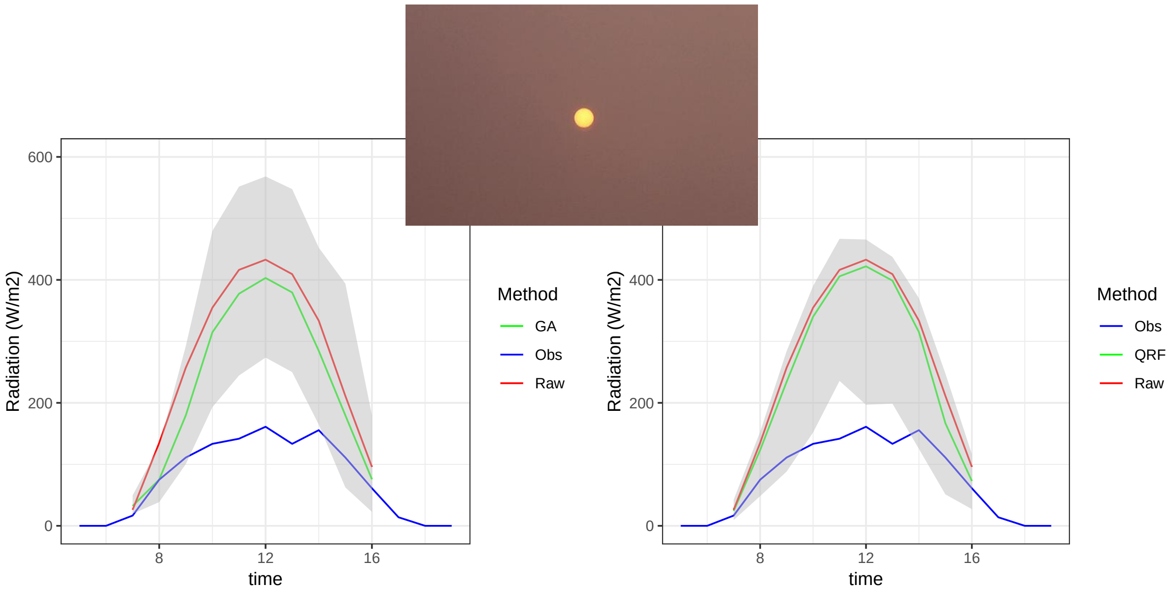


Figure 10: As Figure 7, but for October 17, 2017.

predictors. By comparing Figure 10 and 11 we see that the QRF method performs somewhat worse without the aerosol predictors. The difference in the gray areas is the largest around 12 UTC with a difference between 30 and 40 W/m^2 in the lowest quantile. In this case the aerosol predictors increase the uncertainty towards lower amounts of radiation, which is important, because the observations turned out to be very far away from the median on this day. For the GA method we see similar results with aerosol predictors as without aerosol predictors. This occurs due to the fact that the number of predictors is limited in the GA method and the aerosol predictors are not selected for making predictions.

4.3.2 Continuous ranked probability score

In this and the following subsections, we study how well the methods perform in general and compare them. Therefore we apply scoring metrics on the probabilistic forecasts. We start with the continuous ranked probability score, for which we show the values in Table 6. The scores are calculated over the different seasons for 2 lead times (+12h and +36h) and averaged over the stations. The scores are

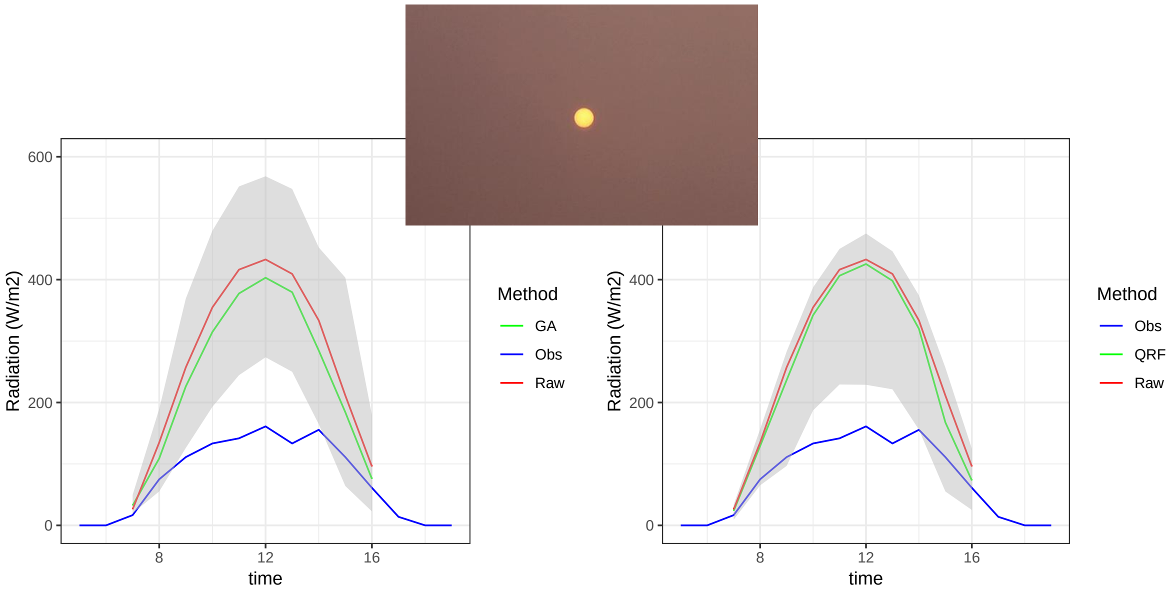


Figure 11: As Figure 10, but now the aerosol predictors from CAMS are not used in the fitting procedure.

calculated for the actual radiation and not for the clear sky index. The climatological CRPS is also listed as a reference.

(a) The CRPS at lead time +12h.

	CLIM	GA	NOtr	QR	QRF	GRF	GBRT	ANN
Winter	66.04	32.99	33.97	**32.30**	33.62	32.55	33.04	34.06
Spring	131.07	76.25	73.94	72.24	70.96	**70.31**	71.60	75.64
Summer	129.46	83.27	82.55	80.44	80.79	**80.16**	80.56	84.82
Autumn	102.76	48.16	48.58	47.40	46.93	**46.79**	46.97	49.19

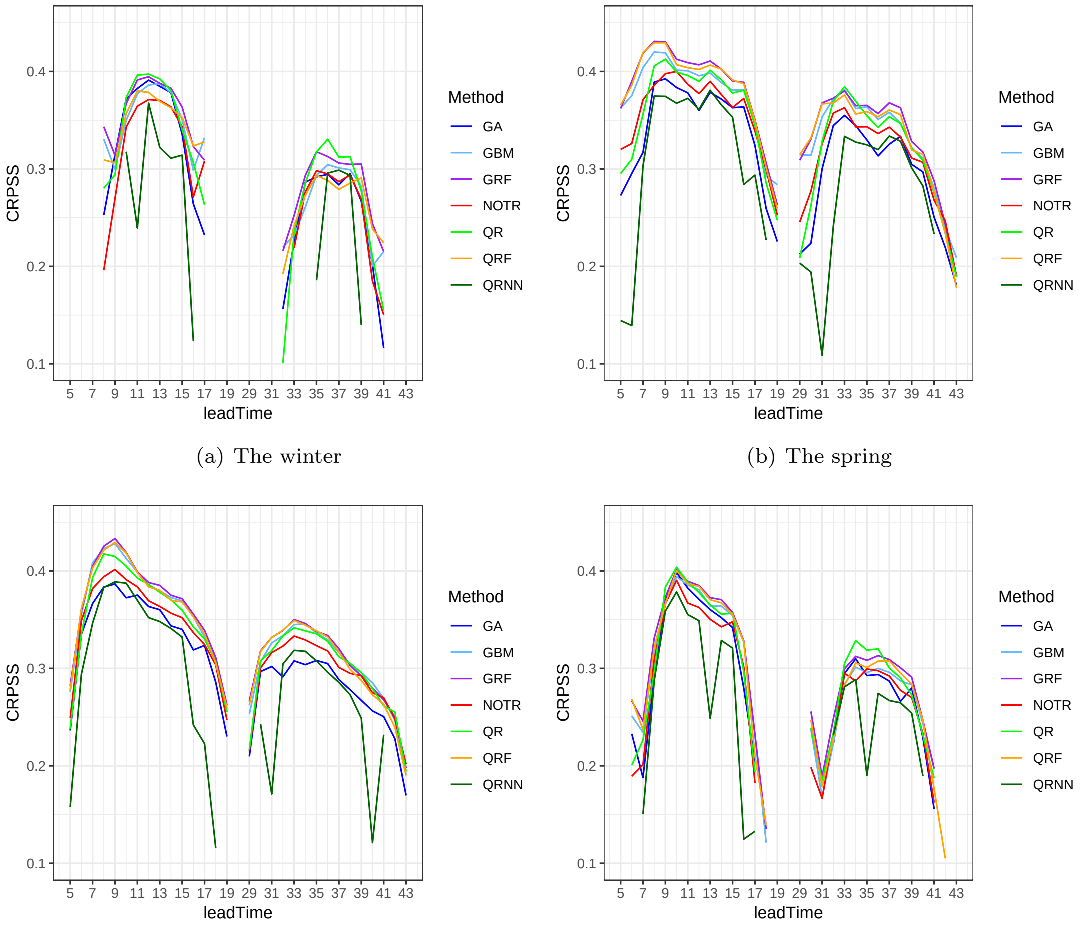
(b) The CRPS at lead time +36h.

	CLIM	GA	NOtr	QR	QRF	GRF	GBRT	ANN
Winter	67.52	39.33	38.72	**36.59**	39.24	37.77	38.32	38.76
Spring	132.52	82.11	79.06	78.21	76.86	**76.48**	77.12	80.92
Summer	129.00	90.79	89.05	87.83	87.29	**87.06**	87.68	91.87
Autumn	102.14	53.65	52.92	**51.24**	52.27	51.80	52.60	54.43

Table 6: The CRPS (Eq. 15) for the different methods calculated over the four seasons at lead times +12h and +36h (averaged over the stations). The lowest (best) score in each row is highlighted. CLIM stands for the sample climatology and the rest of the abbreviations are explained in Section 3.2.

We see that all methods improve greatly with respect to climatology, as expected. The reduction in CRPS is between 50 and 60 W/m^2 for almost all seasons and methods. More interesting is to compare the methods mutually. The best performing method in each case in the table is highlighted and we see, similar to the MAE and RMSE, that in different situation different methods perform best. We see that only the QR and GRF method are highlighted, so those two perform best in terms of the CRPS. However, the CRPS of the rest of the methods is similar to that of the QR and GRF methods, as can be seen in Table 6.

To get a more general picture of the performance of the methods we compute skill scores for the CRPS (denoted by CRPSS). The skill scores are calculated for the clear sky index, so without transforming the predictions first. We show the CRPSS for multiple lead times, multiple stations, the four seasons and the different methods. For the reference forecast we used the sample climatology. We first present figures with the CRPSS for the different methods plotted over the lead times for the four seasons, shown in Figure 12. We see that the CRPSS is higher in the middle of the day than closer to the night. Also the second forecast day shows lower CRPSS than the first forecast day, due to systematic errors increasing with lead time. All methods perform quite similar with maximum skill score values of around 0.45 in the morning in spring and summer. The ANN method performs slightly worse. This is probably due to the complexity of the method. Although we reduced the overfitting as much as possible, a too complex fitting procedure can lead to overfitting and result in worse skill scores. We also compare the seasons and in summer we see that the CRPSS drops significantly during the afternoon and evening, both on day 1 and day 2. This happens due to convection in summer, that can cause cloud formation in the afternoon and evening, which makes the global radiation harder to predict. In the other seasons this trend appears less. In spring and summer the GRF method seems to perform best. In winter and autumn the QR method is mostly performing best.



(a) The winter

(b) The spring

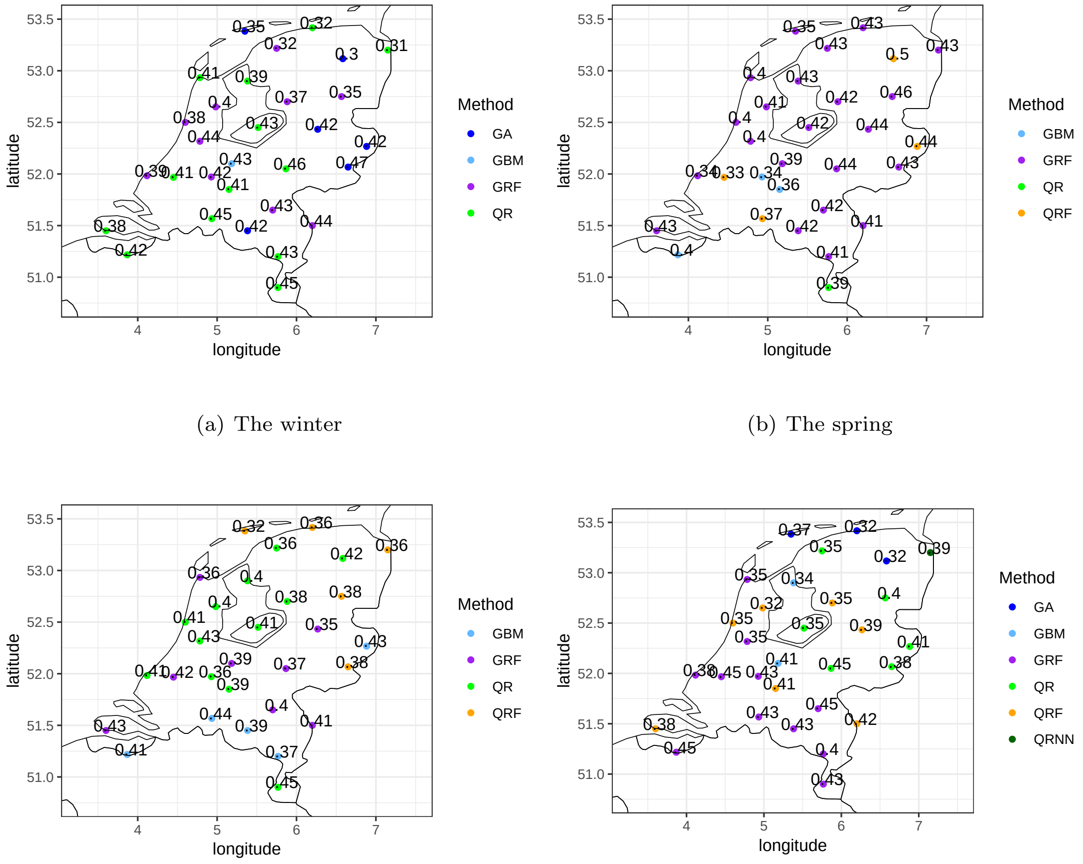
(c) The summer

(d) The autumn

Figure 12: The CRPSS for all lead times calculated over the different seasons and averaged over the stations. The abbreviations of the methods are explained in Section 3.2.

It is also informative to look at the spatial patterns of the CRPSS and therefore we have plot the CRPSS for the different stations in the Netherlands for the lead time +12h in Figure 13. The values in the maps correspond to the best performing method for the stations, with colors indicating what the best performing method is. We see that the best performing methods vary between the seasons and the stations. In winter and spring the GRF method performs best for most stations, but in summer the QR method seems to perform best. For autumn it is less clear, there are many variations between the stations and we cannot conclude that one method performs best for this season. Over all seasons combined we can conclude that the QR and GRF methods perform best, which is in agreement with the results for the comparison over the lead times.

Although the CRPSS values vary between the stations, there is not a clear spatial pattern. We see some seasonal variations in the CRPSS. The CRPSS in winter and autumn are lower for the stations closer to the coast than for stations more inland. This might occur because convection over sea causes more cloud formation over the stations close to the coast and this affects the radiation. In spring and summer there is not a clear spatial pattern in CRPSS. However, because convection occurs mostly in summer for the inland stations, the CRPSS values there are considerably lower in summer than in winter.



(a) The winter

(b) The spring

(c) The summer

(d) The autumn

Figure 13: The CRPSS for all stations calculated over the different seasons for the lead time +12h. Only the maximum value over the methods is shown. The abbreviations of the methods are explained in Section 3.2.

4.3.3 Brier skill score

So far we investigated the effects of time, space and seasons on the CRPSS and from now on we focus on one lead time (+12h) and the whole validation set. We still average the scores over the stations. In this section we investigate how the different methods perform under different weather conditions. Therefore we transform the predictand into a binary predictand by applying a threshold on the continuous predictand and values lower or higher than the threshold are mapped to 1, respectively 0. Because the predictand is the clear sky index, it does not have a daily and yearly cycle and therefore the threshold can be a constant value between 0 and 1. The probabilistic forecasts are transformed to the probabilities of not exceeding the threshold as explained in Section 2.4.3. Using the transformed observations and forecasts we calculate the Brier skill scores using equations (16) and (12) with sample climatology as reference forecast, which is in this case the mean of the transformed observations. We have calculated the Brier skill scores (BSS) for different thresholds $\{\tau = \frac{i}{10}, i = 1, ..., 9\}$ (corresponding to different weather conditions) and plotted the BSS versus the threshold as is shown in Figure 14. We see that the BSS is close to constant for thresholds between 0.3 and 0.8. In that region all methods perform similarly. For thresholds lower than 0.3 (corresponding to fully overcast conditions) the BSS values of all methods decrease. This is due to the fact that with a small threshold almost all observations are mapped to 0. Then the sample climatology also gets close to 0 and the Brier score for the reference forecast gets very low. Then it is hard for the methods to improve the reference Brier score to achieve a positive BSS. The BSS of some of the methods decreases even more than the rest. The BSS of the ANN method starts decreasing already below 0.4, but for the lowest threshold 0.1 the BSS of the GA and QR method have decreased the most (even more than the ANN method). For the highest threshold 0.9 (corresponding to clear sky conditions) we see that there is a bifurcation in the scores. Part of the methods (QRF, GRF, QR and GBRT) keep the BSS at the same value compared to the lower thresholds, but the rest (GA, NOtr and ANN) decrease slightly. This means that the global radiation in clear sky conditions is better forecast by the QRF, GRF, QR and GBRT methods. We also note that all methods have a BSS above zero for all thresholds, meaning that all methods perform better than climatology.

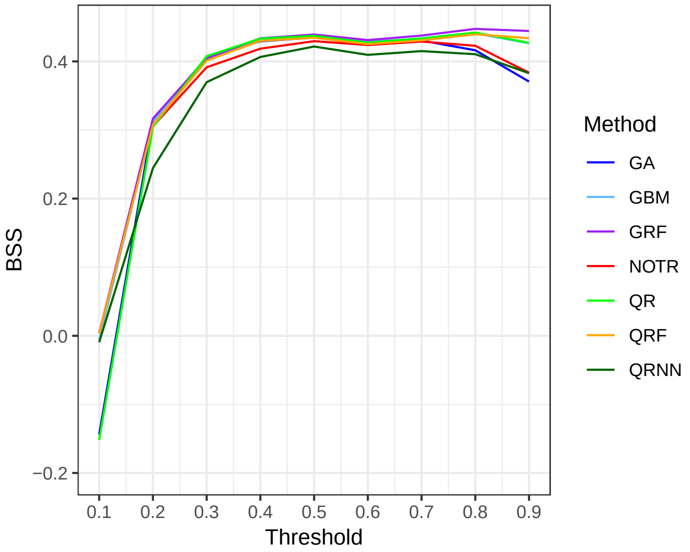


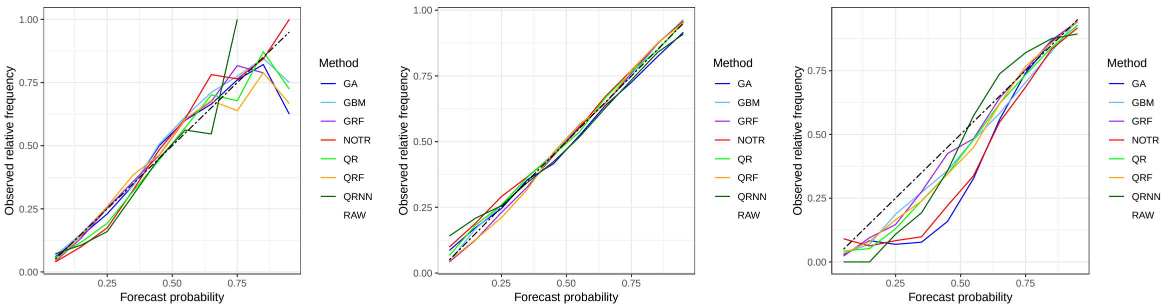
Figure 14: The BSS versus the threshold at lead time +12h, calculated over the complete data set and averaged over the stations. The abbreviations of the methods are explained in Section 3.2.

4.3.4 Reliability diagrams

Reliability diagrams are an interesting way to look at the transformed observations and forecasts. The reliability diagrams for the thresholds 0.2, 0.5 and 0.9 are visualized in Figure 15. The number of bins used for producing the diagrams is 10. As explained in Section 2.4.4 the probabilistic forecasts are more reliable when the forecast probabilities are closer to the observed relative frequencies. This is the case when the plotted lines are closer to the diagonal. For the threshold 0.2 we see that the lines have some spikes for the high forecast probabilities. This occurs because most of the forecast probabilities are low for a low threshold and there are not many cases left for the high forecast probabilities. With only a small number of cases one case can be enough to force the observed relative frequency away from the diagonal. The spikes in the lines indicate that this indeed occurs. For the threshold 0.9 this is not the case, because there are enough clear sky situations in the data set. For the threshold 0.2 all methods are less reliable for the high forecast probabilities except for the NOtr method. For the threshold 0.9 the methods QR, QRF, GRF and GBRT are more reliable than the GA, NOtr and ANN methods. Lastly, for the threshold 0.5 we see that all methods are reliable and there is not a method that performs best. In general we can say that the QRF, GRF, QR and GBRT methods perform best according to the reliability plots.

4.3.5 Potential economic value

An economically relevant way to verify the predictions is to calculate the potential economic value. It is based on a simplified version of the real economic world, but it is still very useful. We look at a range of values for $\tilde{T}$ in maximizing the expected expense for our forecasts and go over a range of cost-loss ratios $\frac{C}{L}$ for calculating the potential economic value. Both $\tilde{T}$ and $\frac{C}{L}$ are defined in Section 2.4.5 and we take the range of values for both as: $\{\frac{i}{100}, i = 1, .., 100\}$. In Figure 16 we have plotted the potential economic value versus the cost-loss ratio for the three thresholds 0.2, 0.5 and 0.9.



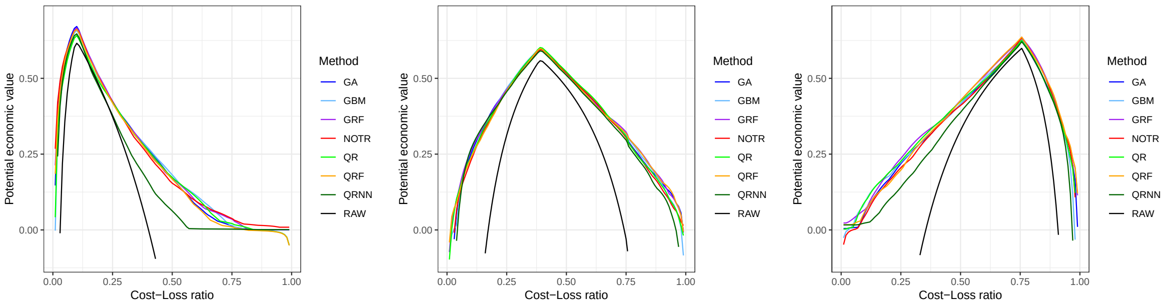
(a) Threshold 0.2

(b) threshold 0.5

(c) Threshold 0.9

Figure 15: The reliability diagrams for different thresholds at lead time +12h, calculated over all dates and stations. The abbreviations of the methods are explained in Section 3.2.

In all plots the highest values are reached when the cost-loss ratio is equal to the observed relative frequency ORF. For the threshold 0.2 the raw forecast has a large potential economic value at the peak, which is at the ORF value of 0.094. However, the potential economic value declines fast for both lower and higher cost-loss ratios than the ORF. This means that it is valuable, but only under certain conditions for the cost-loss ratios, which might not be satisfied in a realistic situation. The other methods increase the potential economic value at the ORF slightly and also have positive values over a much larger range of cost-loss ratios (almost the full range between 0 and 1). The method that gives the maximum potential economic value is the GA method, with a value of 0.67. All methods perform very similar, only the ANN method declines faster when increasing the cost-loss ratio. For the threshold 0.9 the ORF is much higher, namely 0.756, causing the peaks to move to the right part of the plot. Similar to the threshold 0.2, the potential economic values of the methods are higher than the raw forecast and also stay positive over a larger range of cost-loss ratios. This time the QRF method has the highest potential economic value of 0.64 at the ORF. The ANN method is again declining the fastest when decreasing the cost-loss ratio, whereas the rest of the methods have similar performance. For the threshold 0.5 all methods perform similarly (also the ANN method) and improve upon the raw forecast, both in highest potential economic values and in range of positive values. The peaks are now at the ORF of 0.387, with the QR method producing the highest value of 0.6. We conclude that for all thresholds the methods perform similarly. The methods perform however all better than the raw forecast.



(a) Threshold 0.2

(b) threshold 0.5

(c) Threshold 0.9

Figure 16: The potential economic value as a function of the cost-loss ratio for multiple thresholds at lead time +12h, calculated over all dates and stations. RAW stands for the raw forecast and the rest of the abbreviations are explained in Section 3.2.

5 Conclusions

In this thesis we compared parametric and non-parametric regression methods in a probabilistic sense. We produced probabilistic forecasts with the regression methods by means of using the raw forecast (the deterministic forecast of global radiation from HARMONIE) together with other relevant meteorological variables as potential predictors. We saw that all potential predictors were used in the non-parametric methods with the radiation and cloud related predictors being the most important. For the parametric methods only part of the predictors was used, with again the radiation and cloud related predictors as the most important ones. In visualizing the performance of the methods we looked at some interesting cases with different weather conditions. We saw that on the sunny and the fully overcast day the forecast had relatively small uncertainty, whereas on the partly overcast day there was more uncertainty. Also the global radiation on the day with a high amount of aerosols turned out to be hard to predict with a high uncertainty in the forecast. Next we calculated multiple deterministic and probabilistic scoring metrics. We saw that the performance of the methods depends on the time of the day and on the forecast lead time. Also we saw patterns in the geographical location (coast or inland) and between the seasons in the forecast skill. All methods performed better than the raw forecast and had similar performance among each other. Only the artificial neural network method performed worse than the rest. For different scoring metrics we saw different methods performing best. This indicates that it is important to consider multiple scoring metrics and it also illustrates the small differences between most of the methods. According to the CRPSS, quantile regression (QR) and generalized random forests (GRF) performed slightly better than the other methods, consisting of the gamma distribution (GA), truncated normal distribution (NOtr), quantile regression forests (QRF), gradient boosted regression trees (GBRT) and the artificial neural network (ANN). For the Brier skill score the GRF method performed slightly better than the rest. For the reliability diagrams all methods were close to the diagonal and it was less clear which method performed best. For the threshold 0.9 the GA, NOtr and ANN methods were less reliable. We also compared the probabilistic forecasts generated by the methods with the raw forecast. This was first done by comparing the medians of the probabilistic forecasts with the deterministic raw forecast in terms of the mean absolute error and root mean squared error. According to both scoring metrics the medians of the probabilistic forecasts were more accurate than the raw forecast. Also the full probabilistic forecasts were compared with the raw forecast in terms of the potential economic value. This showed that the added uncertainty information from the probabilistic forecasts, in addition to the increased accuracy, led to higher potential economic values and also to positive values over a wider range of cost-loss ratios. This result tells that the probabilistic forecasts are very useful for users of solar energy. The users can make better decisions based on the improved accuracy and information about the uncertainty. Future research could focus on more extended economic models for making a comparison between the methods.

By comparing the non-parametric methods with the parametric methods, we saw that the non-parametric methods have the advantage that they make use of the information from all predictors, whereas the parametric methods only look at a limited amount of predictors. This led in general to better performance of the non-parametric methods (except for the artificial neural network). Also for the case of October 17, 2017, we saw the QRF method taking information from the aerosol predictors in contrast to the GA method. We prefer the non-parametric methods over the parametric ones, due to the fact that the non-parametric methods have more predictive power in general.

Future research could focus on investigating other distributions than the ones tried in this thesis or other non-parametric methods like support vector machines. Furthermore, the effect of more potential predictors (for example the diffuse radiation) could be investigated. Also predictors taken from the (recently started) HARMONIE ensemble prediction system could be used, such as the ensemble mean or ensemble variance of the global radiation forecast, and would likely be better predictors than the deterministic HARMONIE forecasts.

The methods presented here and outlined for the future are all based on techniques of statistical post-processing. A different, physical approach is to improve the NWP models to lead to better predictors. This could be for example improving the representation of clouds and aerosols in the model. Then the forecasts of global radiation will also improve. However, statistical post-processing can still add value on those improved forecasts.

Acknowledgements

We thank Jan Fokke Meirink for providing us with the aerosol forecasts from CAMS, consisting of the aerosol optical depth, Angstrom exponent and the ozone. We thank Wim de Rooy, Jan Barkmeijer, Sander Tijm and Sibbo van der Veen for providing information that is used in this thesis.

References

Almeida, M., Perpin, O., and Narvarte, L. (2015). Pv power forecast using a nonparametric pv model. *Solar Energy*, 115:354–368. https://doi.org/10.1016/j.solener.2015.03.006.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F., and Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111. http://dx.doi.org/10.1016/j.solener.2016.06.069.

Athey, S., Tibshirani, J., and Wager, S. (2016). Generalized random forests. *The Annals of Statistics*. https://arxiv.org/pdf/1610.01271.pdf.

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., De Rooy, W., Gleeson, E., Hansen-Sass, B., Homleid, M., Hortal, M., Ivarsson, K., Lenderink, G., Niemelä, S., Pagh Nielsen, K., Onvlee, J., Rontu, L., Samuelsson, P., Santos Muñoz, D., Subias, A., Tijm, S., Toll, V., Yang, X., and Ødegaard Køltzow, M. (2017). The harmonie-arome model configuration in the aladin-hirlam nwp system. *Monthly weather review*, 145:1919–1935. https://doi.org/10.1175/MWR-D-16-0417.1.

Bracale, A., Caramia, P., Carpinelli, G., Rita Di Fazio, A., and Ferruzzi, G. (2013). A bayesian method for short-term probabilistic forecasting of photovoltaic generation in smart grid operation and control. *Energies*, 6:733–747. https://doi.org/10.3390/en6020733.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. https://link.springer.com/article/10.1023/A:1010933404324.

Cannon, A. (2011). Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & Geosciences*, 37:1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005.

Cannon, A. (2018a). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*. https://www.researchgate.net/publication/326007117_Non-crossing_nonlinear_regression_quantiles_by_monotone_composite_quantile_regression_neural_network_with_application_to_rainfall_extremes.

Cannon, A. (2018b). *qrnn: Quantile regression neural networks*. R package version 2.0.3, https://cran.r-project.org/package=qrnn.

Cervone, G., Clemente-Harding, L., Alessandrini, S., and Delle Monache, L. (2017). Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renewable Energy*, 108:274–286. https://doi.org/10.1016/j.renene.2017.02.052.

David, M., Aguiar Luis, M., and Lauret, P. (2018). Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data. *International Journal of Forecasting*, 34:527–547. https://doi.org/10.1016/j.ijforecast.2018.02.003.

ECMWF (2002). Assimilation techniques (3): 3dvar. https://www.ecmwf.int/sites/default/files/elibrary/2002/16932-assimilation-techniques-3-3dvar.pdf.

Fatemi, S., Kuh, A., and Fripp, M. (2018). Parametric methods for probabilistic forecasting of solar irradiance. *Renewable Energy*, 129:666–676. https://doi.org/10.1016/j.renene.2018.06.022.

Friedman, J. (2001). Greedy function approximation: A gradient boosted machine. *The Annals of Statistics*, 29:1189–1232. https://statweb.stanford.edu/~jhf/ftp/trebst.pdf.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38:367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Future learn. Why the sky is blue. https://www.futurelearn.com/courses/learn-about-weather/0/steps/28848.

KNMI (2018). Hourly meteorological observations. https://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi.

Koenker, R. (2005). *Quantile Regression*. Cambridge U. Press.

Koenker, R. (2018). *quantreg: Quantile Regression.* R package version 5.36, https://CRAN.R-project.org/package=quantreg.

Lorenz, E., Hurka, J., Heinemann, D., and Georg Beyer, H. (2009). Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2:2–10. https://ieeexplore.ieee.org/document/4897348.

Massidda, L. and Marrocu, M. (2018). Quantile regression post-processing of weather forecast for short-term solar power probabilistic forecasting. *Energies*, 11:1–20. https://doi.org/10.3390/en11071763.

McRae, G. (1980). A simple procedure for calculating atmospheric water vapor concentration. *Journal of the Air Pollution Control Association*, 30:394–394. https://doi.org/10.1080/00022470.1980.10464362.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Reseach*, 7:983–999. http://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf.

Meinshausen, N. (2017). *quantregForest: Quantile Regression Forests.* R package version 1.3-7, https://CRAN.R-project.org/package=quantregForest.

Michalsky, J. (1988). The astronomical almanac's algorithm for approximate solar position (19502050). *Solar Energy*, 40:227–235. https://doi.org/10.1016/0038-092X(88)90045-X.

Mohammed, A. and Aung, Z. (2016). Ensemble learning approach for probabilistic forecasting of solar power generation. *Energies*, 9. https://doi.org/10.3390/en9121017.

NHnieuws (2017). In beeld: Zon kleurt oranje door saharastof en bosbranden. https://www.nhnieuws.nl/nieuws/213647/In-beeld-Zon-kleurt-oranje-door-saharastof-en-bosbranden.

Remund, J., Wald, L., Lefvre, M., Ranchin, T., and Page, J. (2003). Worldwide linke turbidity information. https://hal.archives-ouvertes.fr/hal-00465791/document.

Richardson, D. (2000). Predictability and economic value. https://www.ecmwf.int/sites/default/files/elibrary/2003/11922-predictability-and-economic-value.pdf.

Ridgeway, G. (2018). *gbm: Generalized Boosted Regression Models.* R package version 2.1.4, https://CRAN.R-project.org/package=gbm.

Rigby, R. and Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, 54:507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x.

Rigby, R. and Stasinopoulos, D. (2018). *Generalized additive models for location, scale and shape,(with discussion).* R package version 5.1-2, https://cran.r-project.org/package=gamlss.

Rigollier, C., Bauer, O., and Wald, L. (2000). On the clear sky model of the esra european solar radiation atlas with respect to the heliosat method. *Solar Energy*, 68:33–48. https://doi.org/10.1016/S0038-092X(99)00055-9.

Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Miner, L., and Wright, M. (2018). *grf: Generalized Random Forests.* R package version 0.10.1, https://CRAN.R-project.org/package=grf.

Van der Meer, D., Widn, J., and Munkhammar, J. (2018). Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, 81:1484–1512. https://doi.org/10.1016/j.rser.2017.05.212.

Verzijlbergh, R., Heijnen, P., de Roode, S., Los, A., and Jonker, H. (2015). Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications. *Solar Energy*, 118:634–645. https://doi.org/10.1016/j.solener.2015.06.005.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M., Paoli, C., Motte, F., and Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105:569–582. https://doi.org/10.1016/j.renene.2016.12.095.

WeatherWorks (2014). A quick guide to cloud types. https://www.weatherworksinc.com/cloud-types.

Wikipedia (2018). Quantile regression. https://en.wikipedia.org/wiki/Quantile_regression.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences*, volume 100. Academic Press., 3rd edition edition.

A Appendix: Model parameter tuning

In order to achieve the best results for each method, we find the optimal settings in each method and also in the fitting procedure. The model parameters (which we call the "hyperparameters") that define the optimal settings are found by a cross-validation procedure. This gives the optimal settings based on an (mostly) independent testing set. It is not possible to have a fully independent testing set due to the fact that we only have 2 years of data. To make the tuning of the hyperparameters as independent as possible from the final fitting procedure, we choose a different cross-validation procedure for the tuning of the hyperparameters. We do not split the data set into a training and testing set based on separate groups of consecutive dates, but we split the data set randomly over the dates. We still keep all the stations for a specific date in the same set (training or testing).

We start by finding the optimal settings in the fitting procedure. We first look at the effect of averaging the HARMONIE predictors over space and time. We have tried 5 different regions in space (1x1, 3x3, 5x5, 7x7 and 9x9 block of gridpoints), where for each gridpoint closest to a station and for each region we average the HARMONIE predictors over the values in the corresponding region. Each of these regions is centered around the gridpoint. For time averaging, we distinguish between no averaging (i.e the predictors are the values at the lead time) or averaging over 3 lead times, centered around the lead time. The averaging is only applied to the predictors from HARMONIE, because the space and time resolution of the CAMS predictors is too low to make the averaging useful. For the rest of the predictors there is also no need for averaging, because the predictors are (almost) constant on the selected regions in time and space and therefore averaging them would lead to the same value as not averaging them. The predictors from HARMONIE can vary in space and time and therefore averaging can turn out to be useful by averaging out random variations in the predictors. This idea is proven in Figure 17, where we plot the CRPS skill score (CRPSS) against the lead times for the different spatial regions and the time averaging. The skill scores are calculated over the complete independent data set (all dates, because of the cross-validation) and averaged over all stations. The GA method is chosen for the comparison over the spatial regions and the time averaging. We see in Figure 17 that averaging over the 9x9 block of gridpoints and over 3 lead times gives the highest skill scores and therefore the best predictions. For the other post-processing methods or calculating the skill scores only over one of the seasons leads to the same conclusions. Therefore averaging the HARMONIE predictors in space over a 9x9 block and in time over 3 lead times is the best approach and in producing the results these averaged predictors are used.

Next we investigate the difference between fitting the whole data set at once or fitting for the different seasons separately. We still calculate the skill scores over the complete data set for the GA method. The results are shown in Figure 18. We see that the procedure of fitting per season generally gives slightly better skill scores. This is explained by the differences in the weather in different seasons. For example the weather in the summer is different from the weather in the winter, not only in temperature, but also in humidity, cloud cover, etc. Therefore fitting per season allows the fitting procedure to focus more on the possible weather situations in the specific season, while otherwise the fitting procedure has to account for all weather situations that can happen in any of the seasons. This result is also visible for the other methods and also when calculating the skill scores over the seasons separately. Therefore, fitting per season separately is done for producing the results.

A.1 Gamma and truncated normal distributions

The next step is to find the optimal settings in each of the statistical methods. We look at the lead times +8h, +12h, +16h, +32h, +36h, +40h to draw conclusions about the optimal hyperparameter settings. We start with the GA and NOtr method, where we study the effects of two hyperparameters: the number of steps in the stepwise procedure for both the mu and the sigma parameter. Because each step can only add one predictor into the fit, the number of steps is also the maximum number of predictors in the fit. We choose the default for the hyperparameters to be five steps for mu and one step for sigma.

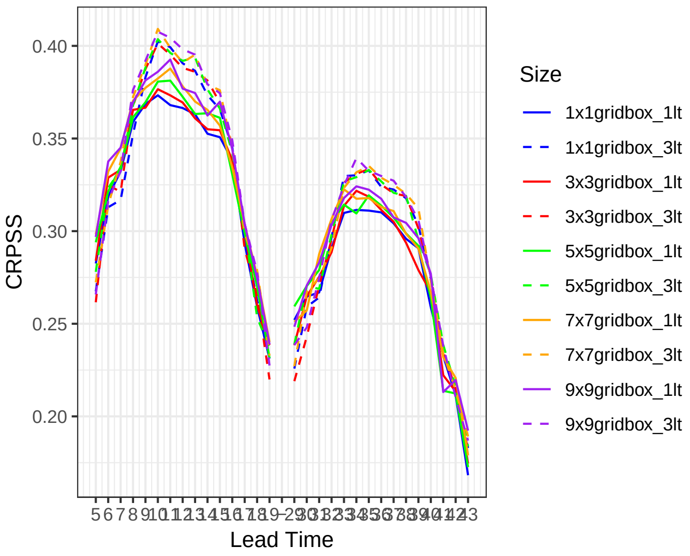


Figure 17: Comparison of CRPSS values between averaging HARMONIE predictors over different spatial regions and over time for the GA method. The CRPSS is calculated over the whole data set and averaged over the stations.

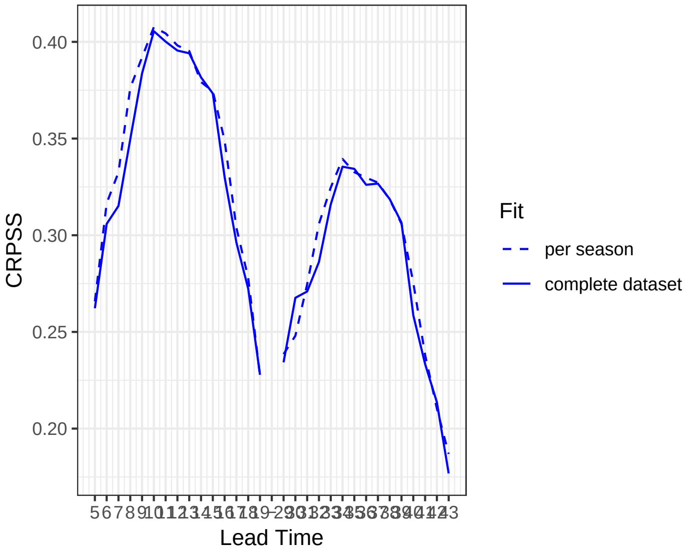
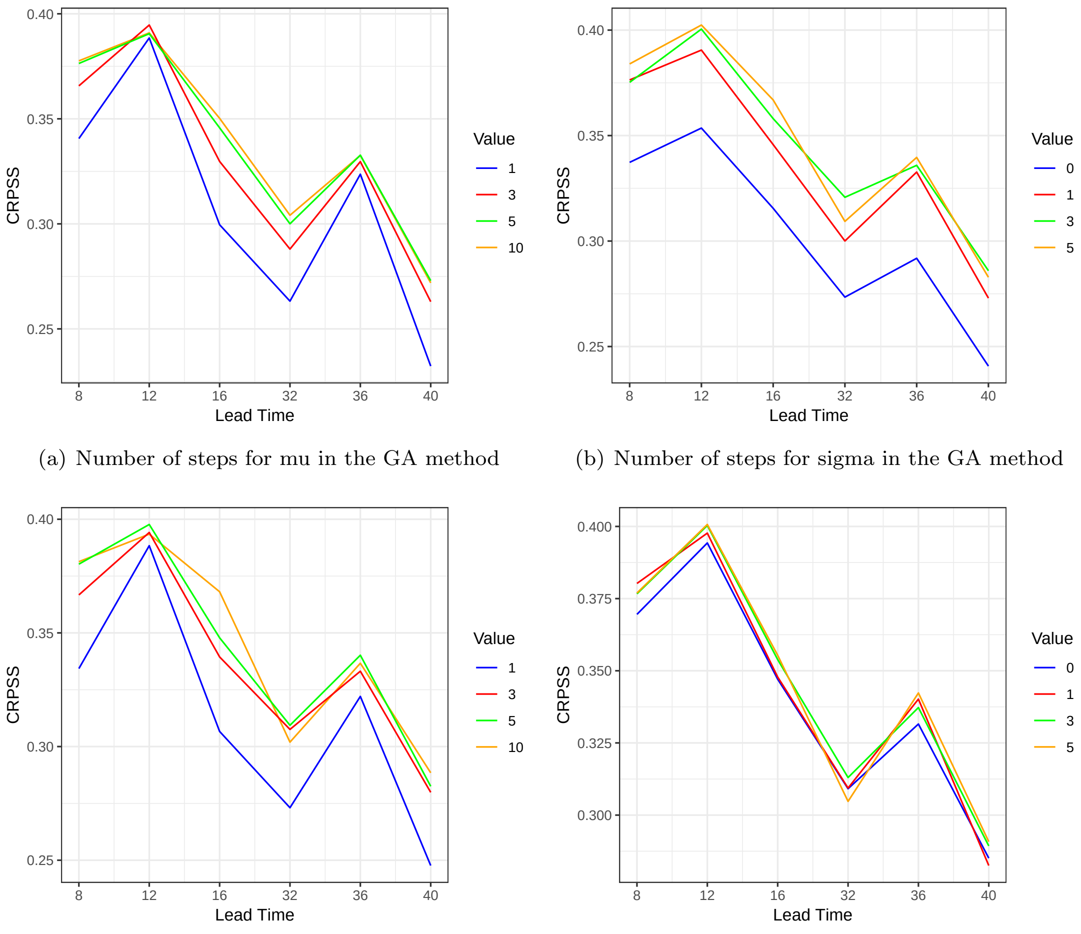


Figure 18: Comparison of CRPSS values between fitting per season or for the whole data set at once for the GA method. The CRPSS is calculated over the whole data set and averaged over the stations.

We change one hyperparameter at a time, while leaving the other at the default. For mu we try 1,3,5 and 10 steps and for sigma we try 0,1,3 and 5 steps. The case with zero steps means that there is only an intercept term for the fit of sigma. In Figure 19 we show the result of varying the hyperparameters for both the GA and the NOtr method. We see that allowing more steps in the stepwise procedure generally leads to better predictions. For the GA method the skill score increases as a function of the number of steps for both mu and sigma. For the NOtr method varying the number of steps for sigma does not have much influence on the skill score. The predictions are more sensitive to changes in mu than to changes in sigma in the NOtr method. Although the skill score values are higher with more steps for mu in both the GA and NOtr method, overfitting is also happening when we apply too many steps. Between 5 and 10 steps for mu there is (almost) no increase in CRPSS anymore and overfitting occurs. This is also seen by looking at the coefficients $\vec{\beta}$ in Equation (1). For example if the cloud cover is one of the predictors that is chosen, then we expect the coefficient for the cloud cover to have a negative sign, because higher values of the cloud cover should lead to lower radiation values. A positive sign for the cloud cover coefficient indicates overfitting. For the number of steps for sigma we see similar skill score values for 1,3 and 5 steps. Also for sigma overfitting can occur when applying too many steps in the stepwise procedure. Therefore we choose 1 step for sigma and 5 steps for mu as the optimal setting in both the GA and NOtr method.



(a) Number of steps for mu in the GA method

(b) Number of steps for sigma in the GA method

(c) Number of steps for mu in the NOtr method

(d) Number of steps for sigma in the NOtr method

Figure 19: The CRPSS for 6 lead times (+8h,+12h,+16h,+32h,+36h,+40h) calculated over the complete data set and averaged over the stations for both the GA (a,b) and NOtr (c,d) method.

A.2 Quantile regression

For the QR method, we also study the effect of changing the number of steps in the stepwise procedure. The number of steps also bounds the number of predictors in the fit. We keep this number of steps the same for all quantiles and as before for the mu parameter, we vary between the values 1,3,5 and 10. The result is shown in Figure 20. We see that allowing more steps generally leads to better skill score values. However there is again the possibility that overfitting occurs. By looking at the coefficients we see that overfitting occurs for 10 steps. For 3 or 5 steps we see that there are a few cases with overfitting, but in general overfitting does not happen. With 5 steps we are able to make better predictions than with 3 steps, because we allow more predictors in the fit. Therefore we choose 5 steps as the optimal setting for the QR method.

A.3 Quantile regression forests

For the QRF method there are more hyperparameters to tune. The method grows a number of trees, for which we try the values 100, 500 and 2000, with the default set to 500 trees. Next, each tree is grown with a stopping criterion in the growing process of a minimal number of observations left in the terminal nodes of the tree. This is referred to as the minimal size of the terminal nodes and the values

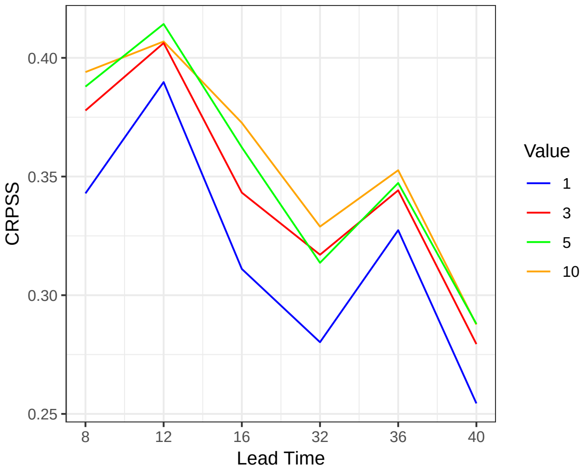


Figure 20: The CRPSS for 6 lead times (+8h,+12h,+16h,+32h,+36h,+40h) and for different number of steps for the QR method (calculated over the complete data set and averaged over the stations).

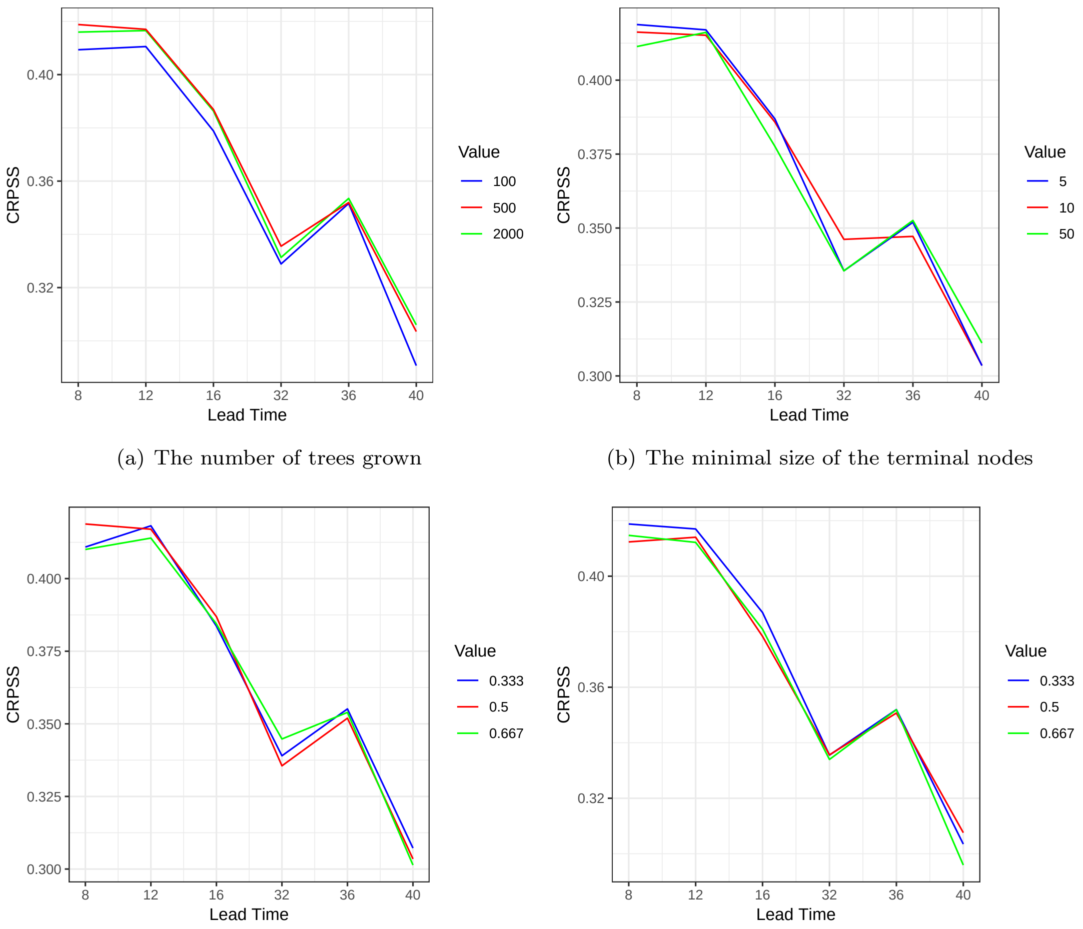
we choose for it are 5, 10, 20 and 50. The default is set at 5. Another hyperparameter to tune is the sampling fraction. Each tree is grown on a different subset of the data and the fraction between the size of the subset and the size of the complete data set is called the sampling fraction. The chosen values for the sampling fraction are $\frac{1}{6}, \frac{1}{2}$ and $\frac{2}{3}$. The default for the sampling fraction is set to $\frac{1}{2}$. The last hyperparameter is the amount of predictors used in making the splits. For each split only part of the predictors are used in finding the optimal split. We choose the fraction of the amount of predictors chosen to the total amount of predictors to vary between the same values as for the sampling fraction: $\frac{1}{3}, \frac{1}{2}$ and $\frac{2}{3}$. The default for this hyperparameter is $\frac{1}{3}$. Both the sampling fraction and the fraction of predictors chosen in growing the trees only tell how large the subset of the data and the subset of the set of predictors are. They do not define the elements in those subsets, because that is decided randomly by a bootstrap sampling procedure. The random component helps in reducing the chance of overfitting. We change the values for the four hyperparameters one at a time, while leaving the others at the default values. The results are shown in Figure 21. We see that the effect of all four hyperparameters on the CRPSS is very limited. Therefore we take the default values for the four hyperparameters as the optimal settings in the QRF method.

A.4 Generalized random forests

The GRF method is almost similar to the QRF method and also has the same hyperparameters to tune: the number of trees, the minimal size of the terminal nodes, the sampling fraction and the fraction of predictors chosen. We choose the same default values for the hyperparameters as for the QRF method and also the same values to tune the hyperparameters. The results for the tuning of the hyperparameters (one at a time, leaving the others at the default) are shown in Figure 22. For all hyperparameters we achieve similar results as for the QRF method and therefore we choose the default values (500 trees, minimal node size of 5, sampling fraction of $\frac{1}{2}$ and fraction of predictors chosen as $\frac{1}{3}$) in the optimal settings for the GRF method.

A.5 Gradient boosted regression trees

For the GBRT method we have five hyperparameters to tune. The method grows a certain amount of trees and each new tree improves the predictions based on the previous tree. The number of trees can be seen as the number of iterations in improving the predictions and we choose the values 100, 500 and 2000 to tune on, with 500 trees as the default. Each tree has a certain depth, which is the number of times new splits have been made on the terminal nodes. This means that for a depth of 1, there was made 1 split. For a depth of 2, there were made 1 + 2 splits. In general for a depth of D, there have been made $\sum_{d=1}^{D} 2^{d-1}$ splits. If the depth increases linearly, then the number of terminal nodes increases exponentially. For the depth values of 1, 5 and 20 are chosen with 5 as the default. We also vary the minimal size of the terminal nodes between 5, 10, 20 and 50, with 5 as the default. Furthermore the



(a) The number of trees grown

(b) The minimal size of the terminal nodes

(c) The sampling fraction

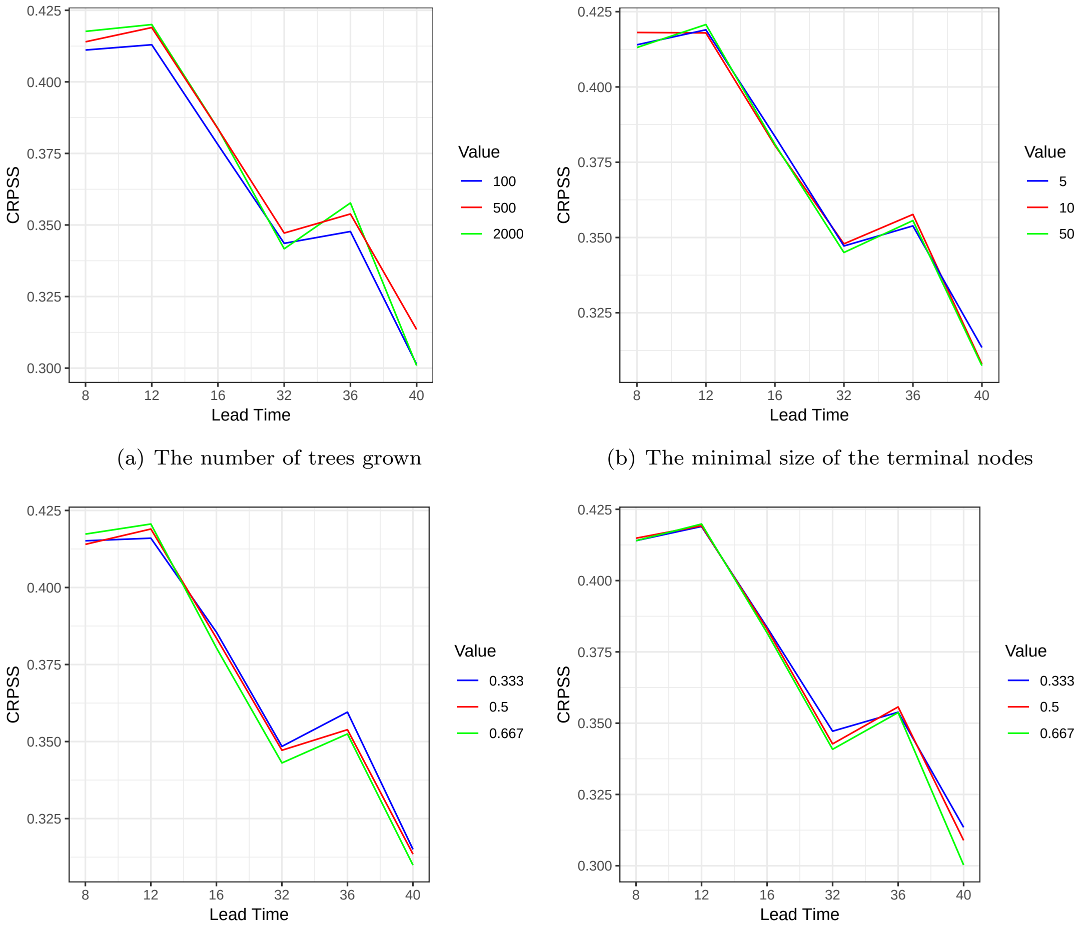
(d) The fraction of predictors chosen

Figure 21: The CRPSS for 6 lead times (+8h,+12h,+16h,+32h,+36h,+40h) for the QRF method, calculated over the complete data set and averaged over the stations.

trees are grown on subsets of the data, so we also have the sampling fraction as a hyperparameter. We take the same values as before: $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$. The default is $\frac{1}{2}$. Lastly, the learning rate is used as a hyperparameter, for which we choose the values 0.01, 0.1, and 1. The default is set to 0.1. The results of changing the hyperparameters one at a time, while leaving the others at the default, are shown in Figure 23.

The CRPSS is not sensitive to the sampling fraction and the minimal size of the terminal nodes. This is because all cases in the data set are selected enough times for all 3 sampling fractions. For the minimal size of the terminal nodes all 3 values lead to enough splits in the trees to achieve good predictions. Also because the depth is limited, in most branches of the trees it is not possible to achieve the minimal size before reaching the maximum depth. We keep the default values for the sampling fraction and minimal size of the terminal nodes in the optimal settings. For the number of trees grown (i.e. the number of iterations), the learning rate and the depth we see that higher values lead to worse CRPSS. This indicates that the procedure is overfitting. With too many iterations, too high learning rate or too large depth, the GBRT method can boost the predictions optimally, leading to predictions very close to the observations. This gives then good skill on the training set, but on the independent testing set it leads to worse skill due to the overfitting on the training set. This is not what we want and therefore we have to make sure that the procedure cannot boost too optimally. On the other hand the procedure has to have a minimum level of boosting to achieve good predictions. Consequently we have to set the number of trees, the learning rate and the depth not too high and not too low. Therefore we try all combinations between them for lead times +12h and +36h and show the resulting skill scores in Table 7:

We see that the combination of the lowest values (100 trees, learning rate of 0.01, depth of 1) and the combination of the highest values (2000 trees, learning rate of 1, depth of 20) both are not optimal



(a) The number of trees grown

(b) The minimal size of the terminal nodes

(c) The sampling fraction

(d) The fraction of predictors chosen

Figure 22: The CRPSS for 6 lead times (+8h,+12h,+16h,+32h,+36h,+40h) for the GRF method, calculated over the complete data set and averaged over the stations.

(a) Lead time +12h

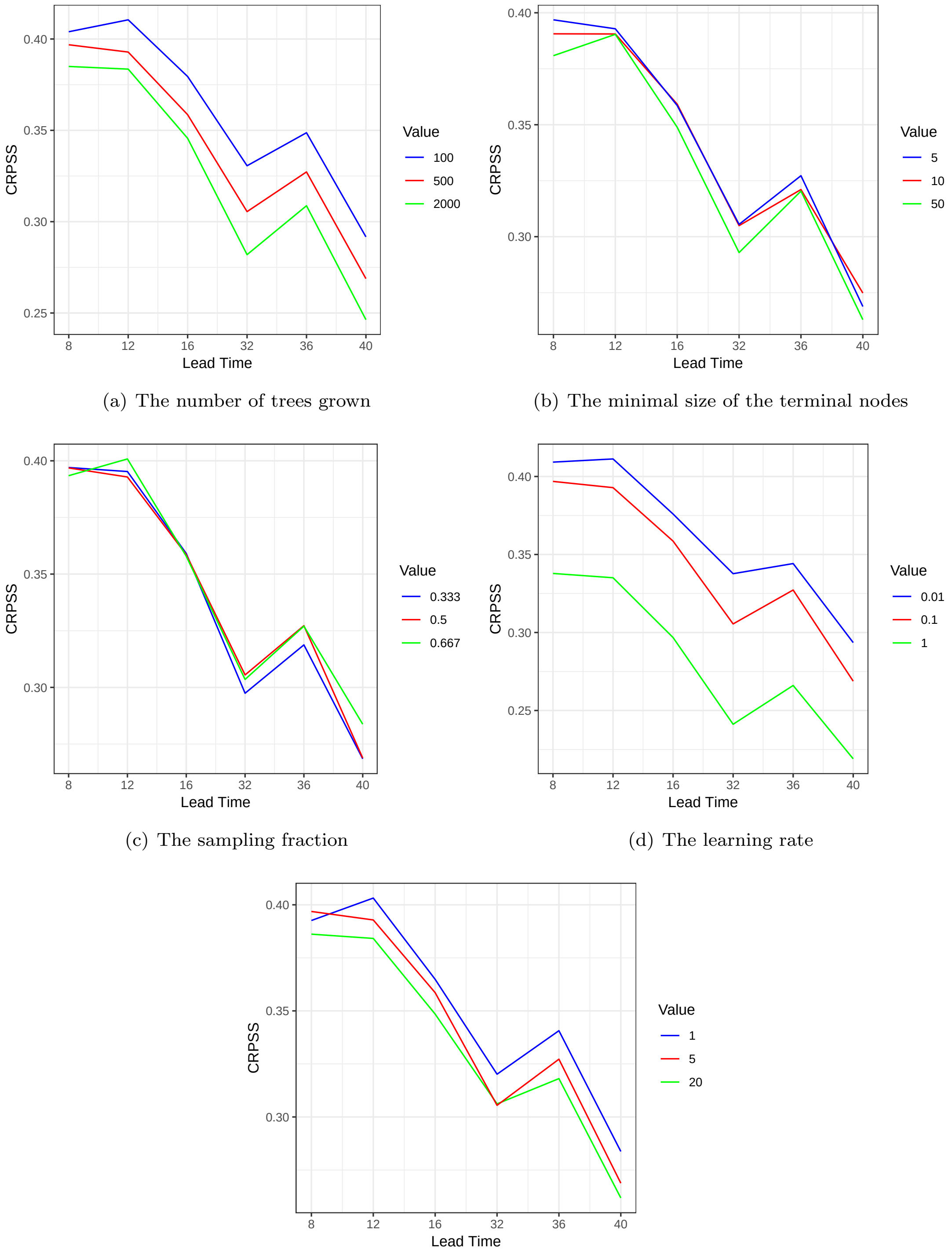
	Trees	100			500			2000		
Learning rate	Depth	1	5	20	1	5	20	1	5	20
0.01		0.238	0.298	0.298	0.403	0.411	0.410	0.411	0.406	0.397
0.1		**0.413**	0.411	0.407	0.403	0.393	0.384	0.388	0.384	0.373
1		0.394	0.370	0.348	0.346	0.335	0.317	0.307	0.312	0.300

(b) lead time +36h

	Trees	100			500			2000		
Learning rate	Depth	1	5	20	1	5	20	1	5	20
0.01		0.208	0.251	0.257	0.344	0.344	0.340	0.342	0.339	0.327
0.1		**0.352**	0.349	0.333	0.341	0.327	0.318	0.304	0.309	0.301
1		0.320	0.311	0.282	0.270	0.266	0.237	0.210	0.242	0.238

Table 7: The CRPSS for lead times +12h and +36h for the GBRT method, calculated over the complete data set and averaged over the stations. The highest values are marked in bold.

in terms of the CRPSS. Actually the best combination is the combination of 100 trees, a learning rate of 0.1 and a depth of 1 for both lead times. Therefore we choose this combination in the optimal settings for the GBRT method.



(a) The number of trees grown

(b) The minimal size of the terminal nodes

(c) The sampling fraction

(d) The learning rate

(e) The depth of the trees

Figure 23: The CRPSS for 6 lead times (+8h,+12h,+16h,+32h,+36h,+40h) for the GBRT method, calculated over the complete data set and averaged over the stations.

A.6 Artificial neural networks

The last method we tune is the ANN method. First we tune the number of iterations in the procedure. We try 2,10 and 100 iterations, with 10 iterations as the default. Next, the number of hidden layers is a hyperparameter that we tune. We try 1,2 and 3 hidden layers, with 3 layers as the default. We also change the number of deep hidden layers between 1,2 and 3 deep hidden layers, with 1 as the default. Lastly, we study the effect of the penalization term in Equation (11) and vary λ between 0,0.01,0.1 and

1, with no penalization ($\lambda = 0$) as the default. The results for tuning the hyperparameters (one at a time, leaving the others at the default), are shown in Figure 24.
For the penalization, we see that increasing λ is affecting the skill scores negatively. Penalizing the weights does not seem to work well. This is probably due to the fact that each weight is penalized in the same way and it would be better to penalize differently for different weights. We keep the default for λ (no penalization) in the optimal settings. For the number of iterations we see that 2 iterations is not yet enough to get the best CRPSS, but 10 or 100 iterations lead to similar CRPSS values. For the number of hidden layers there are more variations between the lead times, but 1 hidden layer works best in general. For the deep hidden layers, we see that 3 hidden layers works best in general. Similar to the GBRT method, the ANN method is a boosting method. To avoid overfitting we should not boost too optimally, but to achieve accurate predictions we also need a minimum level of boosting. Consequently, the optimal settings in the boosting procedure can be between the lowest and the highest values and therefore we tried all combinations between the number of iterations, number of hidden layers and number of deep hidden layers, as is shown in Table 8:

(a) Lead time +12h

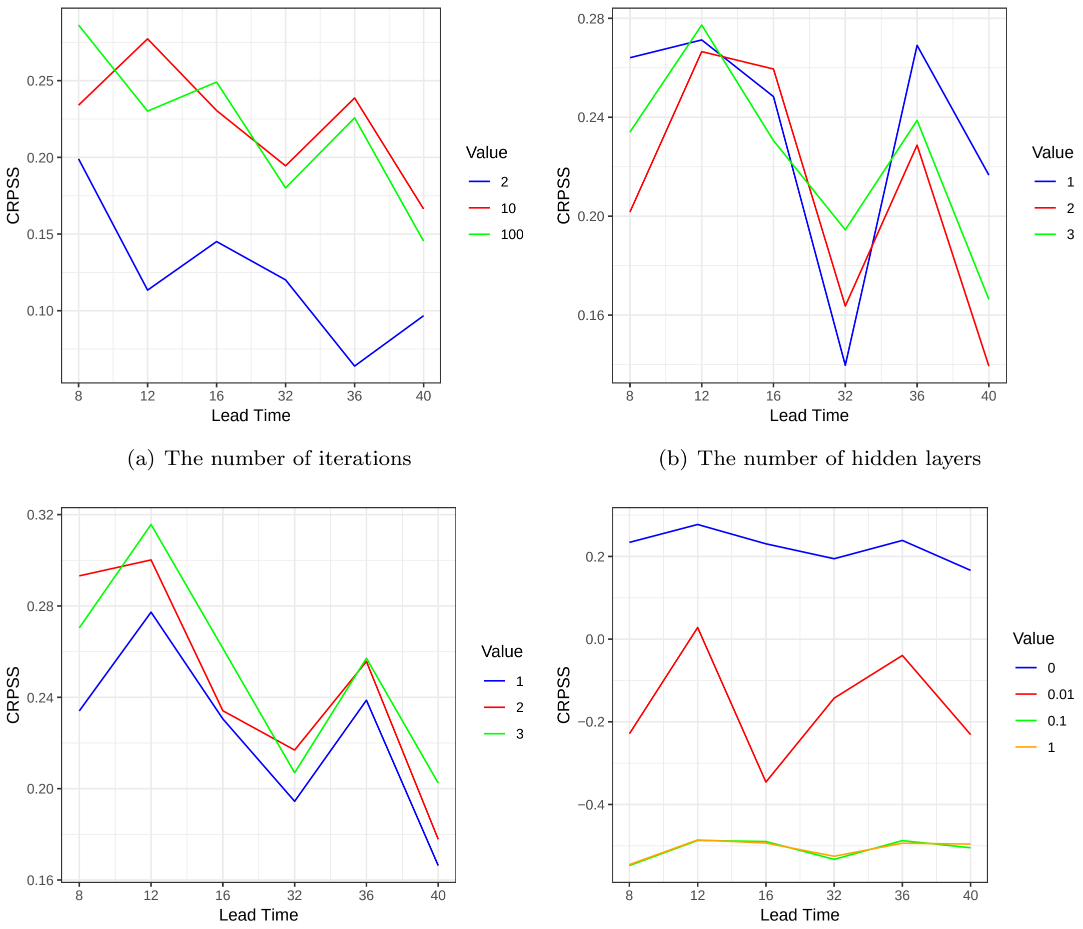
	Hidden layers	1			2			3		
Iterations	Deep hidden layers	1	2	3	1	2	3	1	2	3
2		0.181	0.205	0.206	0.216	0.246	0.244	0.113	0.265	0.259
10		0.271	0.324	**0.388**	0.267	0.342	0.353	0.277	0.300	0.316
100		0.274	0.326	0.368	0.261	0.353	0.375	0.230	0.355	0.367

(b) lead time +36h

	Hidden layers	1			2			3		
Iterations	Deep hidden layers	1	2	3	1	2	3	1	2	3
2		0.150	0.153	0.156	0.166	0.187	0.193	0.064	0.200	0.218
10		0.269	0.299	**0.323**	0.229	0.274	0.274	0.239	0.256	0.257
100		0.229	0.322	0.274	0.221	**0.323**	0.301	0.226	0.305	0.318

Table 8: The CRPSS for lead times +12h and +36h for the ANN method, calculated over the complete data set and averaged over the stations. The highest values are marked in bold.

We see that for both lead times the combination of 10 iterations, 1 hidden layer and 3 deep hidden layers works best. For lead time +36h the combination of 100 iterations, 2 hidden layers and 2 deep hidden layers performs also best. However, this last combination only is best for lead time +36h and also has more computational time due to a higher number of iterations. Therefore, we prefer the first combination of 10 iterations, 1 hidden layer and 3 deep hidden layers and choose that one in the optimal settings for the ANN method.



(a) The number of iterations

(b) The number of hidden layers

(c) The number of deep hidden layers

(d) The penalization term

Figure 24: The CRPSS for 6 lead times (+8h,+12h,+16h,+32h,+36h,+40h) for the ANN method, calculated over the complete data set and averaged over the stations.