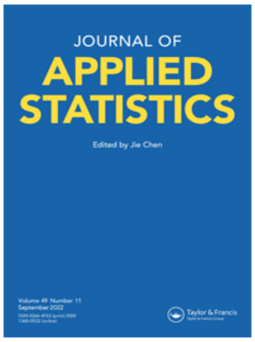


Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Forward variable selection for random forest models

Jasper Velthoen, Juan-Juan Cai & Geurt Jongbloed

To cite this article: Jasper Velthoen, Juan-Juan Cai & Geurt Jongbloed (2022): Forward variable selection for random forest models, Journal of Applied Statistics, DOI: 10.1080/02664763.2022.2095362

To link to this article: https://doi.org/10.1080/02664763.2022.2095362

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2022.2095362

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Forward variable selection for random forest models

Jasper Velthoen [a], Juan-Juan Cai [b] and Geurt Jongbloed [a]

[a]Department of Applied Mathematics, Delft University of Technology, Delft, The Netherlands; [b]Department of Econometrics and Data Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam 1081 HV, The Netherlands

ABSTRACT

Random forest is a popular prediction approach for handling high dimensional covariates. However, it often becomes infeasible to interpret the obtained high dimensional and non-parametric model. Aiming for an interpretable predictive model, we develop a forward variable selection method using the continuous ranked probability score (CRPS) as the loss function. eOur stepwise procedure selects at each step a variable that minimizes the CRPS risk and a stopping criterion for selection is designed based on an estimation of the CRPS risk difference of two consecutive steps. We provide mathematical motivation for our method by proving that in a population sense, the method attains the optimal set. In a simulation study, we compare the performance of our method with an existing variable selection method, for different sample sizes and correlation strength of covariates. Our method is observed to have a much lower false positive rate. We also demonstrate an application of our method to statistical post-processing of daily maximum temperature forecasts in the Netherlands. Our method selects about 10% covariates while retaining the same predictive power.

ARTICLE HISTORY
Received 13 July 2020
Accepted 22 June 2022

KEYWORDS
Random forests; variable selection; CRPS; forward selection; correlated covariates

1. Introduction

In the past decades, random forests [2] have gained traction in many areas of application. Specifically in the last years, random forests have been successfully used for statistical post-processing of weather forecasts, e.g. [19,21] and [18]. A random forest combines several trees, each obtained by recursively making axis-aligned splits in the covariate space until a stopping criterion is reached. The initial algorithm for random forests in [2] provides a good approach for conditional mean regression and classification. Later on, the approach was extended to estimate quantiles by Meinshausen [14] and further improvements were made in [1], which introduced a quantile-based splitting criterion. Due to the results in [1,14], random forests are also used for estimating the conditional quantile function.

These quantile forests have been used in statistical post-processing to obtain probabilistic forecasts, e.g. [18,19,21]. Post-processing is used as a second step in weather forecasting

CONTACT Jasper Velthoen j.j.velthoen@tudelft.nl Department of Applied Mathematics, Delft University of Technology, Mekelweg 4, Delft 2628CD, The Netherlands

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2022.2095362

following a first step of physical modeling, see [11]. This first step entails a numerical weather prediction (NWP) model that uses non-linear partial differential equations of atmospheric flow on a spatial and temporal grid. Together with parametrizations of unresolved physical processes within the grid cells and an estimated initial condition, which is obtained from observational data and a so called first guess (i.e. a forecast for that time based on a previous NWP model run), the NWP model approximates the solution to the partial differential equations. An ensemble prediction system (EPS) adds uncertainty quantification to the NWP model by computing an ensemble of forecasts for perturbed initial conditions and/or the parametrization schemes [11].

Generally there is still a need for bias correction and calibration of numerical weather forecasts, which motivates the second step: statistical post-processing. Historical forecasts together with the corresponding observations are used in post-processing to estimate their statistical relationship. This relationship can then be used in order to calibrate future forecasts.

When post-processing forecasts of a weather phenomenon, a better performance is often attained by adding more information from the NWP models as predictors. For example, [21] showed that the post-processed precipitation forecasts perform substantially better when indices of atmospheric instability from the NWP models are used in modeling the statistical relation. The improvement is due to the fact that the indices of atmospheric instability help to distinguish between different types of precipitation. A full day of drizzle might accumulate to the same amount as a quick shower. However, the distributions of precipitation under these two different weather conditions are very different. Incorporating NWP forecasts of other weather phenomena enables the model to capture such differences.

A natural question is now: ‘Which additional forecasts contain useful information on the phenomenon that one is post-processing?’ The set of potential forecasts to include in the statistical model is generally very large and furthermore they exhibit large correlations. In practice, including too many variables often leads to a decrease in statistical efficiency, and more importantly the model becomes hard to interpret. For a practitioner, it is important to understand which variables play key roles in the statistical model and how they calibrate the EPS forecast. This motivates variable selection procedures in statistical post-processing.

A random forest is generally seen as a method that deals rather well with high dimensional covariates. This property comes from the fact that in the tree fitting algorithm, a random forest chooses the split variables and split points in a greedy way, based on a certain criterion, e.g. the variance. This is often rather effective in the beginning of the tree fitting as many observations are split, but deep down in the tree there are fewer observations which makes the splitting criterion subject to higher variances. Therefore global variable selection methods are considered in the literature to improve statistical efficiency and interpretation of the random forest model.

Variable selection in random forests is mainly done in terms of two types of importance measures. The first type calculates the decrease in impurity of a split made in a tree. In [13] consistency of these measures is shown on fully randomized trees. But in practice in a random forest setting these impurity measures are shown to exhibit biases [17]. The second type is the permutation measure introduced in [2]. This measure computes how much the predictive performance decreases by randomly permuting one single predictor, which breaks the relation between response and the predictor. A popular approach is to perform a backward selection based on the permutation measures, where the model with the best predictive performance is chosen, see e.g. [5,6,9].

Correlation between predictors has a large effect on the permutation importance scores. An initial approach of dealing with this is to consider conditional importance scores, [16]. This has the downside that in some way the conditioning variable has to be chosen. A more precise analysis of the effect of correlation on permutation measures is done in [9], where they conclude that a backward selection is better able to handle correlation between predictors than other strategies incorporating variable importance measures. We show in our simulation study that although the correct variables are often selected by the backward selection, there is no control on the rate of selected noise variables, i.e. the false positives.

In this paper, we propose a new method of selecting variables with random forests. The developed method aims to improve the probabilistic forecasts. Let $Y \in \mathbb{R}$ denote the response variable and $\mathbf{X}$ denote the covariate vector. Given an observed covariate $\mathbf{x}$, a deterministic forecast provides a single value, for instance $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ as the population prediction, whereas a probabilistic forecast offers a conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ as the prediction. The advantage of a probabilistic forecast is that it directly communicates forecast uncertainties that enable the users to make better decisions [3,15]. A commonly used measure of performance for probabilistic forecasts of a scalar variable is the so-called continuous ranked probability score [7]. In our variable selection procedure, we will use this score as the loss function (cf. (1)) to select variables that are informative for the entire conditional distribution instead of just for the conditional mean. The procedure estimates the predictive risk based on the so-called out-of-bag samples (cf. Section 3.2), which is similar to leave-one-out cross validation. At each step, the variable that leads to the smallest CRPS risk is selected and the selection stops if the decrease in CRPS risk is not significantly different from zero. We show by a detailed simulation study that our method controls the false positive rate much better than the backward selection method introduced in [9], even in the presence of high correlations.

The outline of the paper is as follows. In Section 2, we give a detailed description of the mathematical set-up of the variable selection procedure. Then in Section 3, we give a small introduction to random forests and show how the variable selection can be applied to the random forest set-up. A comparison on correlated covariates is given in Section 4 where we compare with an approach selecting variables in a backward selection based on permutation measures. In Section 5, we apply the method to a practical example of post-processing maximum temperature forecasts and compare it to a standard method in post-processing. Finally, we end with a discussion in Section 6.

2. Forward selection

In this section, we describe the mathematical set-up of our forward variable selection method. We provide the intuition of the procedure together with some theoretical motivation of the method in the case of independent covariates.

For now, we consider a pair of random observations $(\mathbf{X}, Y)$, where the response variable $Y \in \mathbb{R}$ and the covariate vector $\mathbf{X} \in \mathbb{R}^d$. Let $J \subset \{1, \ldots, d\}$ denote a set of indices corresponding to the entries of the covariate vector $\mathbf{X}$ and $\mathbf{X}^J$ denote the vector with the entries from $\mathbf{X}$ corresponding to J. Let $F_{Y|\mathbf{X}^J}$ denote the conditional distribution function of Y given $\mathbf{X}^J$. In this section, we work from the population perspective and consider the exact distribution function $F_{Y|\mathbf{X}^J}$ as a forecast of Y. The next section will be concerned with the estimation of the conditional distributions using random forests.

As motivated in Section 1, we are interested in a probabilistic forecast. We use the continuous rank probability score (CRPS) to measure the distance between the degenerate cdf on $\{Y\}$ and $F_{Y|\mathbf{X}^J}$:

$$\mathrm{CRPS}(Y, F_{Y|\mathbf{X}^J}) := \int_{-\infty}^{\infty} \left(I(Y \le z) - F_{Y|\mathbf{X}^J}(z|\mathbf{X}^J)\right)^2 \,\mathrm{d}z. \tag{1}$$

The discrete rank probability score was first introduced in [4]. The application of CRPS was limited due to a lack of tractability of the integral. During the past decades, the CRPS has attracted renewed interest, in particular in the community of meteorology and atmosphere science; see [10,22] among others. As a loss function, the CRPS has two attractive properties. First, it is a proper scoring rule for a large class of distribution functions; see Section 4.2 in [7]. Second, there exists an equivalent expression of the CRPS in terms of the conditional quantile function. Denote the conditional quantile function by $Q_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J)$, where $\tau \in (0,1)$ is a probability level. Then the CRPS can be expressed as

$$\mathrm{CRPS}(Y, F_{Y|\mathbf{X}^J}) = 2\int_0^1 \rho_\tau(Y - Q_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J))\,\mathrm{d}\tau =: \mathrm{CRPS}(Y, Q_{Y|\mathbf{X}^J}), \tag{2}$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ the quantile check function. This equivalent expression of CRPS plays an import role when it comes to the risk estimation. It enables us to evaluate the loss by estimating the conditional quantile via a random forest, detailed in the next section. The equivalence of these two expressions is shown in the appendix. In addition, in Theorem 2.1 below we show that with CRPS as the loss, the procedure is able to retrieve the informative variables correctly.

Corresponding to the CRPS loss, we can now define a risk for the subset of variables corresponding to J as

$$R(J) := \mathbb{E}[\mathrm{CRPS}(Y, F_{Y|\mathbf{X}^J})]. \tag{3}$$

In our approach an ideal variable selection procedure selects the set of variables corresponding to J that minimizes this risk. Define $m_R = \min_{J\subset\{1,\dots,d\}} R(J)$. Then, an optimal set of variables denoted by $\mathbf{X}^{J^*}$ is such that

$$R(J^*) = m_R \quad \text{and} \quad |J^*| = \min\{|J| : R(J) = m_R\} \tag{4}$$

where $|J|$ denotes the cardinality of J. The goal is to identify the smallest model that reaches an optimal risk. This is desirable when it comes to estimating the conditional distribution of Y. It is important to note that J^* is not necessarily unique. For example two collinear covariates X_1 and X_2 both contain the same information on Y, then including any of the two covariates would result in the same expected loss.

In order to obtain J^*, one could evaluate $R(J)$ for all 2^d possible sets, which is often computationally infeasible. Instead we propose a forward variable selection approach as follows. We construct a sequence of length $d+1$ of nested sets J_j for $j = 0, \dots, d$ where

$J_0 = \emptyset$ and

$$J_j = J_{j-1} \cup \left\{ \arg\min_{q \notin J_{j-1}} R\left(J_{j-1} \cup \{q\}\right)\right\}. \tag{5}$$

Our proposed forward selection procedure selects an optimal set J^o such that it is the smallest set attaining the minimum risk among $J_j, j = 0, \dots d$. More precisely,

$$J^o = J_{\min\{j: R(J_j) = \min_{0 \le i \le d} R(J_i)\}}. \tag{6}$$

In the theorem below we show that under the assumption of independent covariates, the sets J^o and J^* coincide.

Theorem 2.1: *Let $X_1, \dots, X_d$ and ϵ be independent random variables. Let $h : \mathbb{R}^{|J^*|+1} \to \mathbb{R}$ be a real valued measurable function and define $Y = h(\mathbf{X}^{J^*}, \epsilon)$, where $J^* \subseteq \{1, \dots, d\}$. Assume that $\mathbb{E}[Y^2] < \infty$, and for any $I \subsetneq J \subseteq J^*$, there exists a set $S \subseteq \mathbb{R}$ with positive Lebesgue measure such that $\mathbb{E}[I(Y \le z)|\mathbf{X}^J]$ is not $\sigma(\mathbf{X}^I)$ measurable for all $z \in S$. Then J^* is the unique subset of* $\{1, \dots, d\}$ *satisfying* (4)*, and $J^0 = J^*$.*

Proof: Let $(\Omega, \mathcal{A}, \mu)$ denote the probability space supporting $X_1, \dots, X_d$ and ϵ. Define the standard inner product on $L^2(\Omega, \mathcal{A}, \mu)$ by $(Z_1, Z_2) = \mathbb{E}(Z_1 Z_2)$, for any random variables Z_1 and Z_2 on $(\Omega, \mathcal{A}, \mu)$. Then $L^2(\Omega, \sigma(X_1, \dots, X_d, \epsilon), \mu)$ becomes a Hilbert space, where the conditional expectation $\mathbb{E}(Z|\mathbf{X}^J)$ is the orthogonal projection of Z onto the closed linear subspace $L^2(\Omega, \sigma(\mathbf{X}^J), \mu)$. Now we have

$$\begin{aligned} R(J) &= \int_{-\infty}^{\infty} \mathbb{E}\left[\left(I(Y \le z) - F_{Y|\mathbf{X}^J}(z|\mathbf{X}^J)\right)^2\right] \mathrm{d}z \\ &= \int_{-\infty}^{\infty} \mathbb{E}\left[\left(I(Y \le z) - \mathbb{E}[I(Y \le z)|\mathbf{X}^J]\right)^2\right] \mathrm{d}z \\ &=: \int_{-\infty}^{\infty} g_J(z)\, \mathrm{d}z. \end{aligned}$$

As the conditional expectation equals the orthogonal projection, for any $z \in \mathbb{R}$,

$$g_J(z) = \min_{G \in \sigma(X^J)} \mathbb{E}[(I(Y \le z) - G)^2]. \tag{7}$$

Therefore, for any $z \in \mathbb{R}$, if $J_1 \subset J_2$, we have

$$g_{J_1}(z) \ge g_{J_2}(z). \tag{8}$$

This implies that $R(J_1) \ge R(J_2)$.

Next, note that if $J_2 = J_1 \cup \{j\}$ and $j \notin J^*$, then for any $z \in \mathbb{R}$,

$$g_{J_1}(z) = g_{J_2}(z). \tag{9}$$

This is because $\mathbb{E}[I(Y \le z)|\mathbf{X}^{J_2}] = \mathbb{E}[I(Y \le z)|\mathbf{X}^{J_1}]$ by the independence of Y and X_j. In this case $R(J_1) = R(J_2)$.

Finally, we show that if $J_2 = J_1 \cup \{j\}$, where $j \in J^*$ then $R(J_1) > R(J_2)$. We prove by contradiction. If not, then $R(J_1) = R(J_2)$, which means in view of (8) that $g_{J_1}(z) = g_{J_2}(z)$, for all $z \in \mathbb{R} \setminus C$, where C has zero Lebesgue measure.

From here we denote $I(Y \leq z)$ by I_z to simplify notation. Expanding the squares and using the tower property of conditional expectation we see that

$$\begin{aligned} g_{J_1}(z) - g_{J_2}(z) &= \mathbb{E}\left[\left(I_z - \mathbb{E}[I_z|\mathbf{X}^{J_1}]\right)^2 - \left(I_z - \mathbb{E}[I_z|\mathbf{X}^{J_2}]\right)^2\right] \\ &= \mathbb{E}\left[-2I_z\mathbb{E}[I_z|\mathbf{X}^{J_1}] + \mathbb{E}[I_z|\mathbf{X}^{J_1}]^2 + 2I_z\mathbb{E}[I_z|\mathbf{X}^{J_2}] - \mathbb{E}[I_z|\mathbf{X}^{J_2}]^2\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[-2I_z\mathbb{E}[I_z|\mathbf{X}^{J_1}] + \mathbb{E}[I_z|\mathbf{X}^{J_1}]^2 + 2I_z\mathbb{E}[I_z|\mathbf{X}^{J_2}] - \mathbb{E}[I_z|\mathbf{X}^{J_2}]^2\right| \mathbf{X}^{J_2}\right]\right] \\ &= \mathbb{E}\left[(\mathbb{E}[I_z|\mathbf{X}^{J_1}] - \mathbb{E}[I_z|\mathbf{X}^{J_2}])^2\right] = 0. \end{aligned}$$

From this we conclude that $\mathbb{E}[I_z|\mathbf{X}^{J_1}] = \mathbb{E}[I_z|\mathbf{X}^{J_2}]$ for all $z \in \mathbb{R} \setminus C$. This implies that $\mathbb{E}[I_z|\mathbf{X}^{J_2}]$ is $\sigma(X_{J_1})$ measurable which contradicts our assumption, hence $R(J_1) > R(J_2)$.

We can now observe that the forward sets are built by adding variables from J^* until all variables of J^* have been added, therefore $J^0 = J^*$. ■

Remark 2.1: Apart from measurability, we do not impose constraints on the function h and the noise term ϵ. It allows a large class of models including additive models, models with interaction terms, etc. When it comes to the estimation, we will use a random forest which is a model-free statistical learning method.

Remark 2.2: The assumption that $\mathbb{E}[I(Y \leq z)|\mathbf{X}^J]$ is not $\sigma(\mathbf{X}^I)$-measurable for any $I \subsetneq J \subseteq J^*$, is needed to prove the uniqueness of J^*. As we know that $R(J^*) = R(J^* \cup \{j\})$ for $j \notin J^*$, there are many sets, which have minimal risk in population sense. The assumption essentially ensures that J^* does contain only indices j such that the function h is not constant for x_j almost everywhere with respect of the distribution of X.

3. Forward selection using random forests

We use a random forest to estimate the conditional distribution function $F_{Y|\mathbf{X}^J}$ and the risk. Now, we make a little excursion to explain the random forest algorithm. We follow the tree construction algorithm proposed in [20] and the extension for quantile estimation from [1]. We choose this approach because it is the only approach that makes splits based on a quantile criterion, additionally in [1] asymptotic normality for the quantile estimates is established.

3.1. Intermezzo: random forests

Denote the data set by $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. A random forest is defined as a collection of trees. Each tree T is obtained by recursively splitting a set of observations by making axis-aligned splits in the covariate space, meaning a split is made on a single covariate value at a time. As a result, every tree induces a partitioning of the covariate space in possibly semi-infinite hyper rectangles. Denote the conditional quantile function by $Q_{Y|\mathbf{X}}(\tau|\cdot)$, where $\tau \in (0, 1)$ denotes a probability level. In this section, we focus on fitting a forest in order to estimate the function $Q_{Y|\mathbf{X}}(\tau|\cdot)$. The estimation procedure for $Q_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J)$ works exactly the same by fitting a forest-based on $\{(\mathbf{X}_1^J, Y_1), \ldots, (\mathbf{X}_n^J, Y_n)\}$.

Recurrent splits are made starting with parent node P, a node in the current partition, creating two child nodes C_1 and C_2, such that $P = C_1 \cup C_2$ and $C_1 \cap C_2 = \emptyset$. This split should be informative with respect to $Q_{Y|\mathbf{X}}(\tau|\cdot)$ and is chosen to maximize,

$$e(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P}(Q_{Y|\mathbf{X}}(\tau|\mathbf{X} \in C_1) - Q_{Y|\mathbf{X}}(\tau|\mathbf{X} \in C_2))^2, \tag{10}$$

where n_P, n_{C_1}, n_{C_2} are the number of observations $\mathbf{X}_i$ in each node. In practice this makes the algorithm very slow as it requires the computation of two quantiles for each possible split. Instead in [1] a relabeling step is proposed and defined as $\mathbb{I}(Y_i > Q_{Y|\mathbf{X}}(\tau|\mathbf{X} \in P))$ for the τ quantile. Now a standard regression split, as used in a standard random forest [2], is made on the labels. This means to maximize the squared difference between the average label in both child nodes.

The trees fitted in [1,20] are called honest trees and are slightly different from the standard structure of tree fitting. A tree is fit by first sub-sampling a set of indices from $\{1, \dots, n\}$ of size $s < < n$ and then randomly splitting this sub-sample in two sets $\mathcal{I}$ and $\mathcal{J}$ both of size $s/2$. Recursive splits of $\mathbb{R}^d$ are then made based on criterion (10), with data points $(\mathbf{X}_i, Y_i) : i \in \mathcal{I}$. The tree becomes honest by removing all the data points indexed by set $\mathcal{I}$ and using only the data points indexed by set $\mathcal{J}$ for estimation of $Q_{Y|\mathbf{X}}(\tau|\mathbf{x})$ for new observations $\mathbf{X}$.

A random forest is then obtained by fitting B trees. Denote by $l_b(\mathbf{X})$ the leaf node of tree b in which $\mathbf{X}$ falls. Then for $1 \leq i \leq n$, the weight for $(\mathbf{X}_i, Y_i)$ induced by the bth tree is given by,

$$w_{i,b}(\mathbf{X}) = \frac{\mathbb{I}(i \in \mathcal{J} \ \& \ \mathbf{X}_i \in l_b(\mathbf{X}))}{\sum_{j \in \mathcal{J}} \mathbb{I}(\mathbf{X}_j \in l_b(\mathbf{X}))}, \tag{11}$$

where $\frac{0}{0} = 0$. The forest weights are obtained by averaging the tree weights over the B trees, $w_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} w_{i,b}(\mathbf{X})$. An estimate of $\hat{Q}_{Y|\mathbf{X}}$ is then given by the locally weighted estimated quantile,

$$\hat{Q}_{Y|\mathbf{X}}(\tau|\mathbf{x}) = \arg\min_\theta \sum_{i=1}^{n} w_i(\mathbf{x})\rho_\tau(Y_i - \theta), \tag{12}$$

with $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ the quantile check function. Note that the structure is similar to kernel regression, but instead of a deterministic kernel with bandwidth h the weights are determined by the data via the forest. Random forests are sometimes called adaptive nearest neighbor estimators for this reason. The tree weights w_i can also be used to obtain the estimator of conditional distribution function:

$$\hat{F}_{Y|\mathbf{X}}(y|\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) I(Y_i \leq y). \tag{13}$$

In the variable selection procedure we aim to select variables that are predictive for the conditional distribution. Therefore, instead of building random forests with respect to a single τ quantile, consider a sequence of quantiles $0 < \tau_1, \dots, \tau_K < 1$. This needs a different type of relabeling than for a single quantile as explained above. They define the relabeling then by,

$$Z_i = \sum_{k=1}^{K} I(Y_i \leq \hat{Q}_{Y|\mathbf{X}}(\tau_k|\mathbf{X} \in P)).$$

The best split is then chosen to maximize the following multi-class classification rule:

$$\hat{e}(C_1, C_2) = \frac{\sum_{k=1}^{K} \left[\sum_{X_i \in C_1} I(Z_i = k)\right]^2}{n_{C_1}} + \frac{\sum_{k=1}^{K} \left[\sum_{X_i \in C_2} I(Z_i = k)\right]^2}{n_{C_2}}.$$

3.2. Estimation of predictive loss

The main quantity in the theoretical framework from Section 2 is the CRPS risk. Replacing the theoretical conditional quantile function by its estimator in (2), we obtain the following targeted loss in the estimation context:

$$\mathrm{CRPS}(Y, \hat{Q}_{Y|\mathbf{X}^J}) = 2\int_0^1 \rho_\tau(Y - \hat{Q}_{Y|\mathbf{X}^J}(\tau|\mathbf{X}^J))\,\mathrm{d}\tau. \tag{14}$$

Here we denote by $\hat{Q}_{Y|\mathbf{X}^J}$ the random forest estimator of the conditional quantile function with respect to the dataset $\{(\mathbf{X}_1^J, Y_1), \ldots, (\mathbf{X}_n^J, Y_n)\}$ and with two arguments, a probability level τ and the covariate vector $\mathbf{X}^J$.

A naive way to estimate the expected loss (that is the expectation of (14)), would be considering

$$\frac{2}{n}\sum_{i=1}^{n}\int_0^1 \rho_\tau(Y_i - \hat{Q}_{Y|\mathbf{X}^J}(\tau|\mathbf{X}_i^J))\,\mathrm{d}\tau.$$

However, this would lead to over-fitting because the training set (data for estimating $Q_{Y|\mathbf{X}^J}$) is the same as the testing set (data for estimating the expectation). This problem can be circumvented by using so called out-of-bag samples as test set.

The out-of-bag samples for the bth tree are defined as the samples that are not used for generating the tree. For each observation $(\mathbf{X}_i^J, Y_i)$, a sub forest $\mathcal{F}_i$ is defined by $\mathcal{F}_i = \{T_b : i \notin (\mathcal{I}_b \cup \mathcal{J}_b)\}$. Namely, this sub forest consists of trees for which $(\mathbf{X}_i^J, Y_i)$ is out-of-bag. Observe that the number of trees in $\mathcal{F}_i$ is random and hence not necessarily the same for all i. The expected number of trees for each sub forest is $B(1 - \frac{s}{n})$.

We use the sub forest $\mathcal{F}_i$ to estimate the conditional quantile function and denote it with $\hat{Q}_{Y|\mathbf{X}}^{\mathcal{F}_i}(\tau|\mathbf{X}^J)$. Since the trees in sub forest $\mathcal{F}_i$ do not use observation $(\mathbf{X}_i^J, Y_i)$, we use this quantile estimator to evaluate the CRPS loss for $(\mathbf{X}_i^J, Y_i)$. Doing this for all observations, we obtain the estimated CRPS risk given by,

$$\hat{R}(J) := \frac{2}{n}\sum_{i=1}^{n}\int_0^1 \rho_\tau(Y_i - \hat{Q}_{Y|\mathbf{X}^J}^{\mathcal{F}_i}(\tau|\mathbf{X}_i^J))\,\mathrm{d}\tau. \tag{15}$$

In the sequel, we write $\hat{Q}_{Y|\mathbf{X}^J}^{\mathcal{F}_i}(\tau|\mathbf{X}_i^J) = \hat{Q}^{\mathcal{F}_i}(\tau|\mathbf{X}_i^J)$ for simplicity.

This out-of-bag procedure for estimating risk has similarities with leave-one-out cross validation. For validating the quantile of the ith observations we use all trees which do not use the ith observation. The difference is that sub forests have in expectation the same size, but not exactly. Computationally the out-of-bag sample approach is also much faster compared to leave-one-out cross validation. Note that a tree has $n-s$ out-of-bag samples and hence the tree is used in $n-s$ sub-forests. On the other hand leave one out cross validation does not reuse trees and estimates a new forest for each element in the summation of (15).

3.3. One step forward

The forward variable selection sequentially adds variables such that the predictive loss is minimized. We here explain how each step is performed. Let d denote the total number of covariates and d' denote the number of signal variables, that is $d' = |J^*|$. In order for the estimation procedure to work, we need a sparsity assumption: $d' << d$.

Recall that for an index set J, the estimated risk $\hat{R}(J)$ is given by (15). Suppose that we have selected the first $j-1$ variables with indices in $\hat{J}_{j-1}$. Then the jth variable $X_{\hat{i}_j}$ is selected based on

$$\hat{i}_j = \arg\min_{q \notin \hat{J}_{j-1}} \hat{R}(\hat{J}_{j-1} \cup \{q\}). \tag{16}$$

and $\hat{J}_j = \hat{J}_{j-1} \cup \{\hat{i}_j\}$. The procedure of a single step forward is detailed in Algorithm 1.

Algorithm 1: A forward step

```
Result: $\hat{i}$, $\hat{R}(J \cup \{q\}) : q \notin J$
Define data $(Y_1, \mathbf{X}_1), \ldots (Y_n, \mathbf{X}_n)$;
Define set $J \subset \{1, \ldots p\}$;
for $q \notin J$ do
    construct a forest with $(Y, \mathbf{X}^{J \cup \{q\}})$;
    Calculate $\hat{R}(J \cup \{q\})$;
end
$\hat{i} = \arg\min_q \hat{R}(J \cup \{q\})$
```

3.4. Stopping criterion

Motivated by the result in Theorem 2.1, we stop selecting variables when there is no further decrease in CRPS risk. From the proof of Theorem 2.1, adding variables that are not in J^* does not have an effect on the CRPS risk. In practice, where we are working with finite samples, additional covariates decrease in fact the statistical efficiency of the random forest which leads to higher CRPS values. Because of the random component in the forest procedure, different forests will have different risk. In general this can be avoided by fitting an enormous number of trees to reduce the random component, but in practice this is infeasible due to the computational load. Instead we make use of the randomness to design a stopping criterion. Ideally, the procedure stops at jth step if $R(\hat{J}_{j-1}) - R(\hat{J}_j) = 0$ and continues if the difference is positive. However, the difference is unknown and we estimate it via the already fitted forests at jth and $(j+1)$th steps. More precisely, we estimate this difference by $\hat{R}(\hat{J}_{j-1} \cup \{q\}) - \hat{R}(\hat{J}_j \cup \{q\})$, where $q \notin \hat{J}_j$. Note that $\hat{R}(\hat{J}_{j-1} \cup \{q\})$ is computed at the jth step for identifying $\hat{i}_j$ and $\hat{R}(\hat{J}_j \cup \{q\})$ at the $(j+1)$th step for identifying $\hat{i}_{j+1}$. So, this estimation does *not* require any extra forest fitting.

Now suppose that the variable added at jth step is a noise variable. Then for any q, $R(\hat{J}_{j-1} \cup \{q\}) - R(\hat{J}_j \cup \{q\}) = 0$ and the estimate $\hat{R}(\hat{J}_{j-1} \cup \{q\}) - \hat{R}(\hat{J}_j \cup \{q\})$ fluctuates around zero due to the randomness in the estimation. Therefore,

$$\mathbb{I}(\hat{R}(\hat{J}_{j-1} \cup \{q\}) - \hat{R}(\hat{J}_j \cup \{q\}) > 0) \overset{\text{asym.}}{\sim} \text{Bern}(0.5).$$

Combining all the information for $q \notin \hat{J}_j$, we define

$$W_j := \sum_{q \notin \hat{J}_j} \mathbb{I}(\hat{R}(\hat{J}_{j-1} \cup \{q\}) - \hat{R}(\hat{J}_j \cup \{q\}) > 0). \tag{17}$$

Then approximately $W_j \sim \text{Bin}(M_j, 0.5)$, where $M_j = d - |\hat{J}_j|$ under the assumption that the jth selected variable is a noise variable. The procedure stops at jth step if the following stopping criterion is met: $W_j < C^j_{1-\alpha}$, where $C^j_{1-\alpha}$ is the $1-\alpha$ quantile of $\text{Bin}(M_j, 0.5)$ and $\alpha \in (0,1)$ is a pre-specified level. The mathematical motivation for the stopping criterion is provided in the appendix.

In practice, the integration in (15) is numerically approximated. Let $\tau_t = \frac{t}{k+1}, t = 1, \ldots, k$, where k is a pre-specified integer. The estimated risk $\hat{R}(J)$ in (15) is approximated by

$$\hat{R}(J) = \frac{2}{k} \sum_{t=1}^{k} \rho_{\tau_t}(Y_i - \hat{Q}_{Y|\mathbf{X}}(\tau_t|\mathbf{X})). \tag{18}$$

The complete procedure is given in Algorithm 2.

Remark 3.1: This stopping criterion works well when M_j (number of variables not been selected yet) is sufficiently large. This is more or less assured by our sparsity assumption of $d' << d$. In other words, this algorithm assumes that there are sufficiently amount of noise variables. This is also the situation when variable selection is interesting.

For the situation when there are only a few or no noise variables, we need some adjustment in the algorithm. The algorithm stays the same until M_j becomes small, that is, most variables are selected. Then stopping criterion in (17) will automatically stop with selecting new variables. Because for a small M_j, one has $C^j_{1-\alpha} = M_j$ due to the discreteness of a binomial distribution and the stopping criterion $W_j < C^j_{1-\alpha}$ is met. For instance, if $M_j = 4$ then $C^j_{0.95} = 4$. To adjust the algorithm to accommodate this issue, we suggest to estimate the risk differently. For a sufficiently large L, we fit $2L$ random forests to estimate the risk $R(\hat{J}_{j-1})$ and $R(\hat{J}_j)$, each with L forest respectively. The adjusted stopping criterion is defined as

$$\tilde{W}_j := \sum_{l=1}^{L} \mathbb{I}(\hat{R}_l(\hat{J}_{j-1}) - \hat{R}_l(\hat{J}_j) > 0),$$

where $\hat{R}_l(\hat{J}_{j-1})$ is the estimate of $R(\hat{J}_{j-1})$ based lth random forest. The procedure stops if $\tilde{W}_j < \tilde{C}_{1-\alpha}$, where $\tilde{C}_{1-\alpha}$ is the $1-\alpha$ quantile of $\text{Bin}(L, 0.5)$. Note that this is computationally much more demanding due to the fact that $2L$ more random forest models are computed at each step.

Algorithm 2: Forward variable selection

```
Result: $J^o$
Set data $(Y_1, \mathbf{X}_1), \dots (Y_n, \mathbf{X}_n)$;
Set $j = 1, J_0 = \emptyset, \alpha$;
Calculate $J_1$ with Algorithm 1 using $J = J_0$;
repeat
    $j = j + 1$;
    Calculate $J_j$ with Algorithm 1 using $J = J_{j-1}$;
    Calculate $W_j$ from Equation 17;
until $W_j \leq C^j_{1-\alpha}$;
$J^o = J_{j-1}$;
```

4. Comparison based on simulation

In this section we assess the performance of our variable selection procedure for correlated explanatory variables. We will compare with the backward selection based on a permutation measure with a mean squared error criterion proposed in [9]; details of the method are stated later in this section. We compare with this method as it is currently the only method that deals with selection with correlated predictors for random forests and we will refer to it as the backMSE method. For the comparison we simulate data from the following model,

$$Y = \mu(\mathbf{X}) + \sigma(\mathbf{X})\epsilon, \tag{19}$$

where ϵ follows a standard normal distribution and independent of this, $\mathbf{X} \in \mathbb{R}^{25}$ follows a multivariate normal distribution. For the covariance structure of $\mathbf{X}$ we split up the covariates into blocks $I_l = \{(l-1)*5+1, \dots, (l-1)*5+5\}$ for $l = 1, \dots, 5$. The covariance matrix of $\mathbf{X}$ is then given by,

$$\mathrm{Cov}(X_j, X_i) = \begin{cases} 1, & \text{if } i = j; \\ \rho, & \text{if } i, j \in I_l \quad \text{for the same } l; \\ \eta, & \text{otherwise.} \end{cases} \tag{20}$$

Observe that the value of ρ controls the correlation strength within each block and η controls the strength between blocks. We generate data from four simulation models with sample size $n \in \{500, 1000, 2500\}$. In the first three models, the covariates are simulated with a block independent structure where signal variables are independent of each other. Precisely for Models 1-3, we set $\rho \in \{0, 0.4, 0.8\}$ and $\eta = 0$. The μ and σ for each model are specified below. The sparsity assumption made in Section 3.3 is satisfied within these models as $d = 25$ and $d' = 3$.

Model 1: $\mu(\mathbf{X}) = X_1 + \frac{X_6}{2} + \frac{X_{11}}{4}$ and $\sigma(\mathbf{X}) = 1$.
Model 2: $\mu(\mathbf{X}) = X_1$ and $\sigma(\mathbf{X}) = \exp(\frac{X_6}{2} + \frac{X_{11}}{3})$.
Model 3: $\mu(\mathbf{X}) = \begin{cases} X_6^2, & \text{if } X_1 \geq 0, \\ -X_6, & \text{if } X_1 < 0, \end{cases}$ and $\sigma(\mathbf{X}) = \begin{cases} 2, & \text{if } X_{11} \geq 0, \\ 1, & \text{if } X_{11} < 0. \end{cases}$

For the first three models, with different values of ρ we investigate the effect of different correlation strengths between the noise variables and the signal variables on the selection procedure. The difference between the three models is in how the signal variables influence the response. In the first model Y only depends on covariates through its mean function. In the second model the dependence is both on the main and the variance. The third model considers discontinuous covariate dependence on both mean and variance.

We consider a fourth model which introduces correlation cross blocks, so the signal variables are dependent now and the covariance matrix of predictors is no longer block diagonal. The dependence between Y and the signal variables is the same as Model 1.

Model 4: $\mu(\mathbf{X}) = X_1 + \frac{X_6}{2} + \frac{X_{11}}{4}$ and $\sigma(\mathbf{X}) = 1$, where $\rho = 0.8$ and $\eta \in \{0.2, 0.4, 0.6\}$.

The backMSE method evaluates the relevance of a covariate by its permutation importance measure, which is defined as

$$I(X_j) = \mathbb{E}\left[(Y - \mathbb{E}(Y|\mathbf{X}_{(j)}))^2\right] - \mathbb{E}\left[(Y - \mathbb{E}(Y|\mathbf{X}))^2\right],$$

where $\mathbf{X}_{(j)} = (X_1, \dots, X_j', \dots, X_d)$ such that $X_j' =^d X_j$ and X_j' is independent of Y and of the other covariates. A large score of $I(X_j)$ indicates that covariate X_j is important. The method randomly permutes the values of X_j to mimic a random sample of X_j'. An estimator of $I(X_j)$ using out of bag samples is given in (2.1) in [9].

In [9] it is shown that the order of the permutation importance measures can not be naturally interpreted in the presence of correlation between the covariates, as variables that are correlated share their importance. As a result, the importance of the important variables is lower than it should be. The backMSE deals with this problem by iteratively removing the least important variable and refitting the model and calculating the importance scores. This process is repeated until no variables are left. The optimal model is then chosen as the model that minimizes the out-of-bag mean squared error. Why this works is easily seen with two highly correlated informative variables. Initially they do not seem important because they share their importance, but by removing one the importance is not shared any more. The left over variable shows the true importance and will therefore be in the selected set.

It is recommended in [9] to compute several forests and take averages to stabilize the variable importance scores and the error estimates. We compute for each step 20 forests where each forest contains 2000 trees. For this method, we follow the standard forest algorithm from [2], fitting trees based on bootstrap samples of size n, *mtry* is set to the default value for regression $p/3$ and taking a minimum leaf size of 5.

For our method we also take fixed parameters with sub-sampling fraction $s = 0.5$, a minimum node size of 1, $mtry = p$ and 1000 trees. We have tested the influence of these tuning parameters on several simulation models and the results are rather robust to different choices. Our selection model adds variables one at a time and stops when additional variables do not increase performance. As the model is therefore often small it makes sense to not over randomize by setting *mtry* to smaller than p. We advise to choose a small s for large datasets in order to reduce computation time.

For each model we simulate 100 data sets. The results are summarized in Figure 1. We show for each model, sample size and covariate dependence structure the average number of correctly selected signal variables in blue, the average number of incorrectly selected noise variables in red and the total number of signal variables with the black line. From this the average number of missed signal variables is visualized as the difference between the black line and the height of the blue bar.

For the first model we see that the backMSE method retrieves more signal variables than the forward selection for low sample sizes and that as the sample size grows the forward variable selection also recovers all signal variables. A large difference is seen in the number of noise variables that are selected. The forward selection performs much better in this than the backMSE, which systematically selects noise variables and tends to even select more as the sample size increases. This phenomenon is also visible for Models 2 and 3 as seen in Figure 1(b, c). For these two models where the variance is dependent on covariates, the CRPS criterion clearly has an edge over the backMSE that selects variables based on the mean squared error and therefore has a hard time selecting these variables. For Model 4 similar results are visible to the Model 1 with correlation strength $\rho = 0.8$, indicating that increased correlation between signal variables does not affect the overall quality of the result.

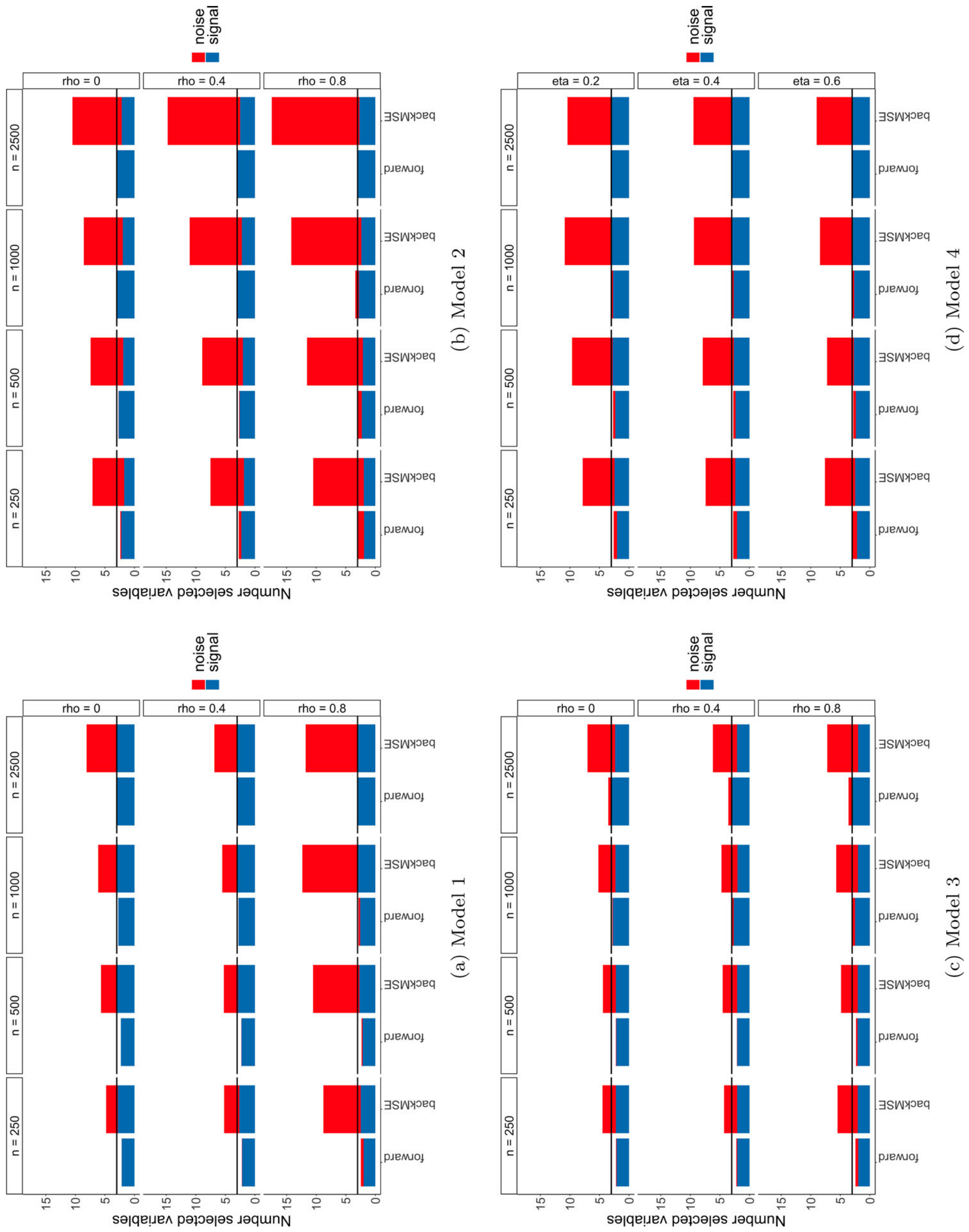
The reason why the backMSE selects many noise variables is two-fold. First the backMSE method selects the optimal set based on a predictive mean squared error criterion. This approach does not account for the inherent variable selection within the random forest, where at each node the split that reduces the variance the most is chosen. As a result the random forest is able to ignore noise variables partially. In practice this means that in an out-of-bag performance measure the addition of a single noise variable cannot be detected. Therefore the variables that are selected will not be the smallest set, but instead a set with maximum number of noise variables maintain the lowest performance. Secondly, the backMSE does not adequately deal with the correlation. For example in Model 1 with $\rho = 0.8$ all variables $X_1, \dots, X_5$ have higher variable importance compared to X_6, which means that if X_6 is in the model, so are $X_2, \dots, X_5$.

Thanks to our testing approach, a small number of noise variables is selected with the forward selection. Using the randomness induced by the random forest, our testing procedure selects a variable that leads to a significant reduction of the predictive loss. The significance level naturally controls the number of selected noise variables by the nature of the testing procedure. We have set the significance level to 5% for all simulations in the paper.

4.1. Variable selection in the setting $d > n$

We end this section with a simulation in a situation where $d > n$, i.e. when there are more variables than observations, and show that our method is able to deal with these high dimensional problems. The dimension of the covariate space is set to $d = 150$ and sample size $n = 100$. The dependence between the covariates has a similar block independence structure as for Model 1, but the number of blocks is now 30. Dependence between the covariates and the response is defined by (19) with functions,

$$\mu(\mathbf{X}) = X_1 + X_6 + X_{11}$$
$$\sigma(\mathbf{X}) = 1.$$



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 1. Average number of selected variables over 100 simulations. Total number of variables is the height of the bar blue for signals variables and red for noise variables. Bars from left to right correspond to the sample sizes and top to bottom corresponds to different levels of correlation, except for Model 4 where top to bottom corresponds to minimum correlation strength between blocks. The horizontal line coincides the total number of signal variables in the model.

Note that also here the sparsity assumption is satisfied with $d = 150$ and $d' = 3$. The experiment is repeated for 100 times and the results for our forward method and the backMSE method are displayed in Figure 2.

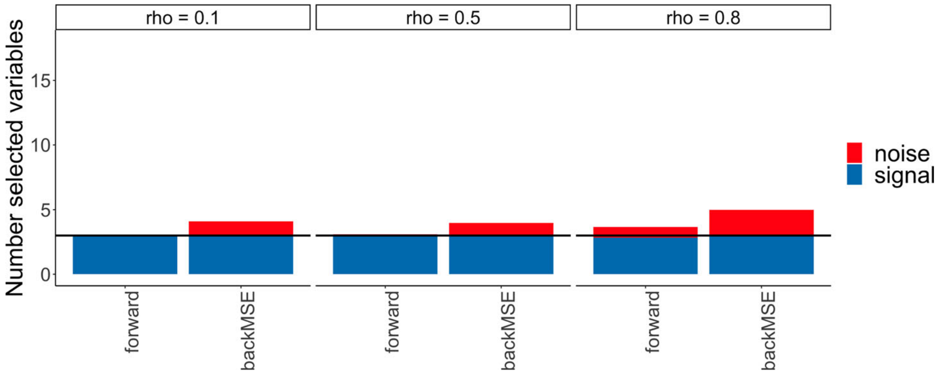


Figure 2. Average number of selected variables over 100 simulations for $n = 100$ and $d = 150$. The blue bars correspond to the signal variable the red bars to the noise variable. The covariate dependence structure from left to right changes with the indicated correlation strength (see label). The horizontal black line coincides with the total number of signal variables in the model.

It can be observed that our method is able to select all signal variables all of the time, only when correlation strength is 0.8, more variables are selected. The backward method again systematically selects more noise variables for all correlation strengths, which is consistent with the results observed in the setting with $d < n$.

5. Post-Processing maximum temperature forecasts

There are substantial risks related to extremely high temperatures. Consecutive days of high temperatures, i.e. heat waves, lead to higher mortality, especially among older people. Besides high temperatures can cause train rails to expand and thereby potentially disrupt the train system. Additionally, in the absence of rain they likely cause severe droughts as seen in 2018 in The Netherlands, which has had large consequences for nature areas and agriculture. The Royal Netherlands Meteorological Institute (KNMI) issues alarms for persistent warm weather. To design a good alarm system it is essential to have good quality weather forecasts. One of the most used ensemble models, the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble model, has a negative bias in the maximum temperature forecast. As an illustration, Figure 3 shows the forecast bias for data observed at weather station de Bilt where KNMI is located. For accurate forecasts, this bias needs to be corrected for. This can be easily done by estimating the linear relation between the forecasts and the observations. Although this quickly improves the maximum temperature forecast, this leaves unused a vast amount of forecast data for other weather types. We will show that using a wide range of potential covariates, the maximum temperature forecasts are improved further than by a simple bias correction. By performing the variable selection we then also investigate in more detail what effect different covariates have on the forecast distribution estimated using the random forest model.

We use maximum temperature observed at seven stations spread across The Netherlands, namely Den Helder, Schiphol, De Bilt, Eelde, Twente, Vlissingen and Maastricht

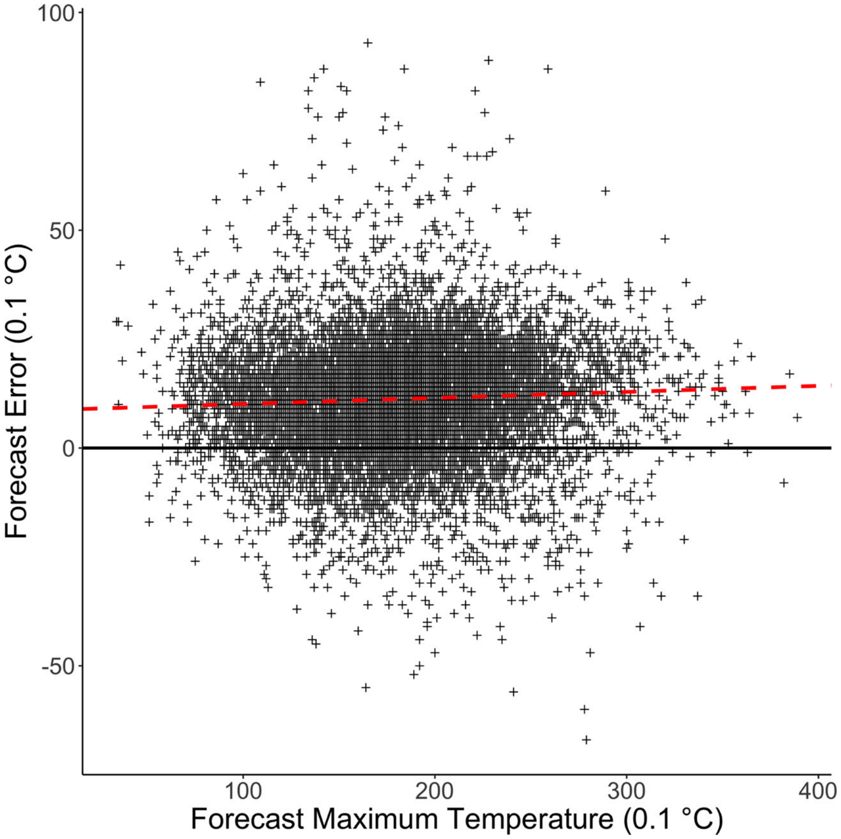


Figure 3. Scatter plot of error of the ECMWF high resolution deterministic run maximum temperature forecast against that deterministic forecast in the warm half year for the years 2011-2019 at De Bilt. The black line indicates a zero error and the red dashed line is the linear regression of the data points.

(http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi). The focus is on high temperatures, hence we consider only observations from mid-April until mid-October. In total, we look at 9 years of data ranging from 2011 to 2019.

As covariates we use the output of the ECMWF model, which contains a 51 member ensemble and a higher resolution deterministic run. These forecasts are initiated two times a day, at 00 UTC and at 12 UTC, but here we use only forecasts of the latter run. We define the lead time of the forecast as the time difference between the start of the day for which the forecast is valid and the initiation time of the forecast. For this analysis we will consider forecasts with lead times equal to $36 + 24k$ hours for $k = 0, 1, 2, 3, 4, 5$. The ensemble contains 51 exchangeable members and in order to use them we compute a set of summary statistics from the ensemble. These summary statistics are the mean, standard deviation, quantiles and number of ensemble members exceeding a pre-specified threshold. For the quantiles in our application we choose the 25%, 50% and 75 % quantiles. Thresholds are chosen as to extract different types of information from the ensemble relative to the weather phenomenon itself. For cloud cover we use three thresholds, 20 %, 50 % and 80 % of cloud cover to create variables measuring probabilities of a few to no clouds, partly clouded weather and clouded weather.

Apart from the forecasts for maximum temperature and cloud cover, we consider other covariates including forecasts for daily average temperature at 2 m, dew point temperature, minimum temperature, daily average wind speed and daily accumulated precipitation. For long lead times, predictability of these typical weather phenomena decreases, but the range of predictability of for example flow pattern at 500 hPa extends much further. Therefore the first three principal components flow pattern at 500 hPa over Europe are also used as predictors [12]. Note that these covariates are the same for each station.

For the response variable we consider the forecast error, which we obtain by subtracting the deterministic forecast run from the observed maximum temperature. By doing so, the seasonality in the temperature is largely reduced. In Figure 3, the forecast error is clearly visible as the distance between the red linear regression line and the x-axis is rather large. Additionally it is clear that the spread of error changes as a function of the deterministic forecast. A possible explanation is that there still remain seasonality effects that are not taken care of by a simple linear effect. Therefore, also the sine and cosine of the day of the year with a period of one year and half a year are included as two predictors. In total this gives us 71 covariates. For a given lead time an observation on a given day is denoted by $(Y, \mathbf{X})$, with Y the error of the deterministic run and $\mathbf{X}$ the 71 dimensional covariate vector.

In what follows, we will explore 3 methods, quantile random forests as in [1] with all variables, quantile random forests with variables selected by our forward variable selection and Non-homogeneous Gaussian Regression (NGR) [8]. This third method is known in the meteorology literature as an EMOS (Ensemble Model Output Statistics) method and is used as a standard approach in post-processing. The NGR method assumes the data follow a Gaussian model,

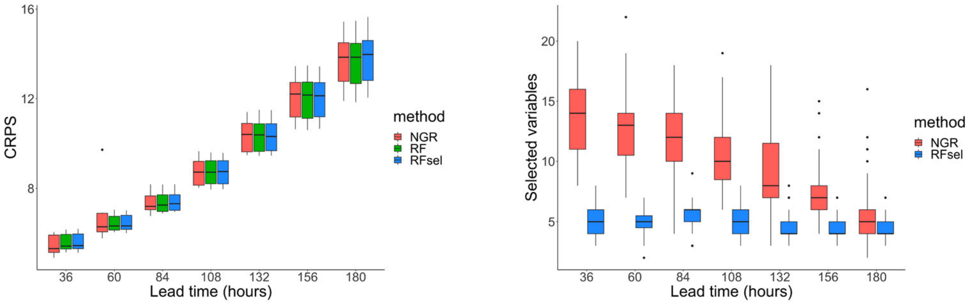
$$Y|\mathbf{X} = \mathbf{x} \sim N\left(\mathbf{x}^T\beta, \exp(\mathbf{x}^T\gamma)\right).$$

The parameters β and γ are then estimated by maximum likelihood. For this model, we select variables based on the Bayesian Information Criterion (BIC) by a forward and backward stepwise approach.

For each station and lead time, we fit a separate model. The models are estimated with a 9-fold cross validation, each time leaving out a single year. In Figure 4(a) the CRPS risk is shown as a function of lead time, where the box-plots contain the CRPS risk for all stations. Then in Figure 4(b) the number of selected variables is shown for our method and NGR, where we leave out the random forest with all 71 variables.

Based on the CRPS, all methods perform comparably. This is also confirmed by other verification measures such as reliability diagrams, quantile reliability diagrams and probability integral transform histograms, which are not shown in this paper. A selection of these diagrams is shown in the appendix. The interesting part comes from the number of selected variables. Our method selects a small portion (less than 10%) of covariates, substantially less than NGR. We investigate this further by considering which variables are selected. The result is visualized in Figure 5. For each lead time, the color indicates the frequency of a covariate being selected by 63 estimated models (7 locations and 9 cross-validation sets per location). An extremely important variable would be selected all 63 times.

Yellow boxes correspond to a few variables that are always selected. But the number of light blue boxes is much smaller for our method compared to NGR. From this we conclude that our method selects fewer variables and it also selects similar variables for different



(a) CPRS

(b) Selected variables

Figure 4. (Left panel) the CRPS risk against leadtime where the boxplots contain the CRPS risk for each station. (Right panel) the number of selected variables against leadtime.

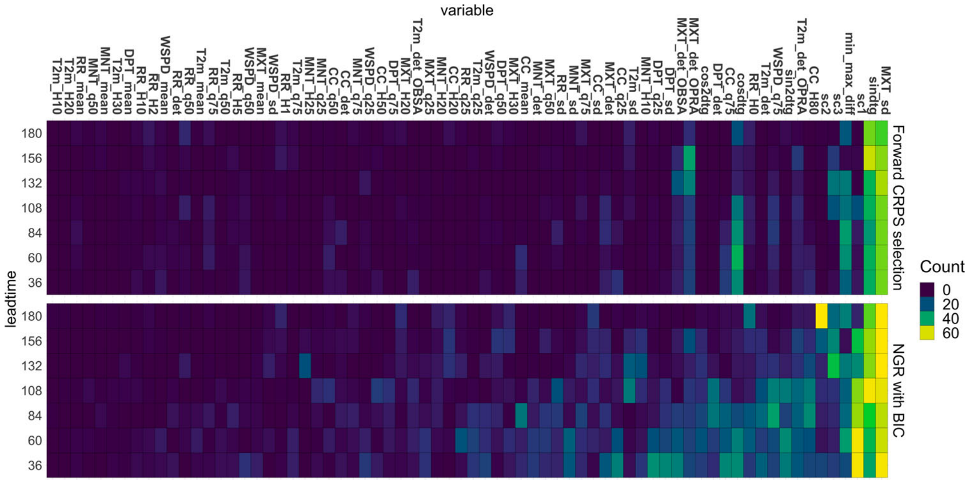


Figure 5. Selected variables for each lead time. All cross validations and all stations have been aggregated where the maximum number of times a variable can be selected is 63.

stations. This suggests that our variable selection method is more robust compared to the NGR method for short lead times, where a diverse set of variables is selected.

The main variables that our method selects are the sine of the day of the year, the standard deviation of the ensemble forecast and variables related to cloud cover. Since our procedure typically selects a small set of variables, it is then feasible to interpret the estimated model. For instance, to investigate how a selected covariate, say X_j influences the forecast distribution of Y, one can compare the conditional distribution of Y given different values of X_j while the other covariates denoted by $\mathbf{X}_{(-j)}$ are kept the same. We consider Y, the forecast error at de Bilt with lead time 36 h and X_j the cloud cover, which is the number of ensemble members with cloud cover exceeding 50%. The values of other covariates are fixed the same as the data of 31-05-2018 at De Bilt, denoted by $\mathbf{X}_{(-j)} = \mathbf{x}^*_{(-j)}$. Figure 6 shows the conditional density of Y given $(X_j = c, \mathbf{X}_{(-j)} = \mathbf{x}^*_{(-j)})$, where different colors indicate three different values of c. Note that all 51 ensemble members exceed 50% cloud cover. As shown in the lower panel of Figure 6, cloud cover clearly has an effect based on

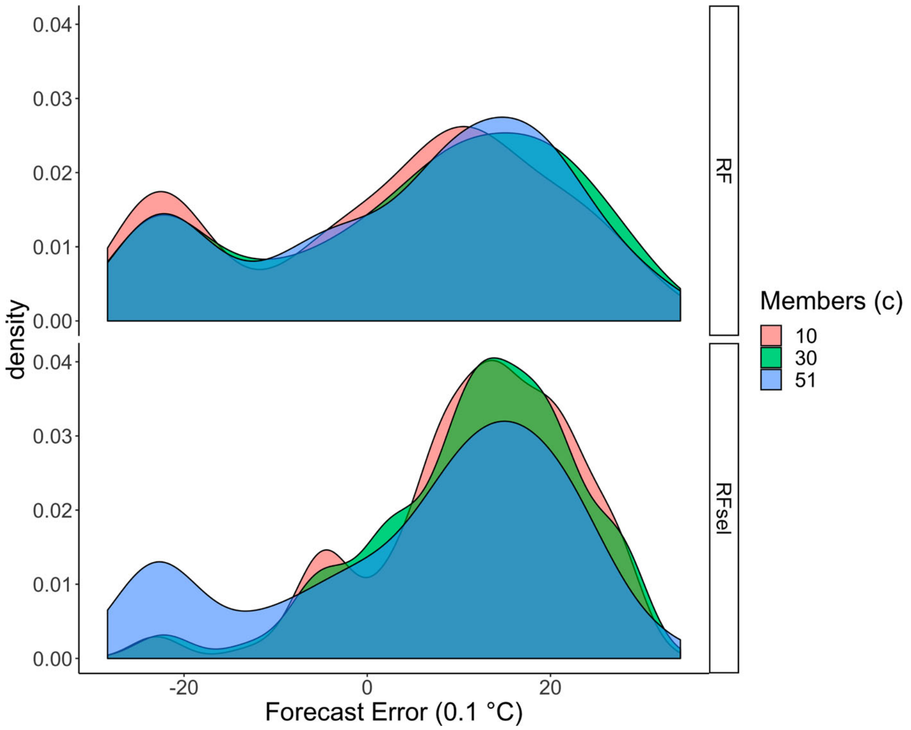


Figure 6. The conditonal distribution of the forecast error, based on estimated models for De Bilt with lead time 36 h. Different colors indicate different values of cloud cover while the values of other covariates are fixed to be the same as that for 31-05-2018 at de Bilt. Top figure for the random forest model with all variables and the bottom figure for the random forest model with the selection of variables.

the estimation of our method: $c = 51$ yields a bimodal distribution while $c = 10$ leads to a unimodal distribution. This suggests that in this configuration, higher cloud cover implies a higher chance for a negative forecast error (left mode in the plot). However, the distributions obtained by random forest (without variable selection) are very similar; see the upper panel of the figure. This is because that there are other covariates correlated to cloud cover, and these covariates still indicate that there is a high cloud cover even when the number of ensemble members exceeding 50% could cover is set to 10. In other words, changing the value of a single variable in a random forest with many correlated covariates is *not* interpretable. Such a random forest model fails to capture the effect of a signal variable.

6. Summary and discussion

In this paper, we have proposed a general framework for a forward variable selection with respect to a loss function. We show in population sense that under an independence assumption between covariates and by choosing the continuous ranked probability score as loss function that the forward selected variables form the correct set with respect to the CRPS risk functional. Applying the method in a random forest set-up, we show that the out-of-bag samples can be efficiently used to asses predictive performance. The main difficulty in the procedure is determining the stopping time, that is when selecting more variables does not add in predictive performance. Due to randomness and the inherent greedy variable selection procedure in the random forest algorithm this can not be determined by the calculated predictive performance. Instead in a single forward selection step we use the predictive performance of each possible set to construct a test to detect increasing predictive performance. The procedure then stops a null hypothesis of non increasing predictive performance can not be rejected. We show that this test is consistent.

With a simple simulation study we show that our variable selection method, compared to a backward selection based on a permutation importance measure, is more capable of discriminating between signal variables and noise variables. This improvement is shown for various sample sizes and correlations between the covariates. Additionally we show that similar results are obtained in a high dimensional setting where $d > n$.

In an application on post-processing maximum temperature, our method shows consistency in the number of selected variables and in the variables being selected over several stations. Moreover, our method selects less than 10 % of the covariates and still attains the same predictive power as the quantile random forest with all covariates. Further, it is easier to interpret our resulting model, due to the largely reduced number of covariates. Without variable selection, it is hardly possible to analyze the effect of a single covariate in a random forest model when it is heavily correlated to other covariates. In our data example, in the presence of thick cloud cover, our random forest model indicates that there is a higher risk of over forecasting (lower panel of Figure 6) instead of under-forecasting which was indicated by Figure 3.

There are two interesting directions for future research. First, the theoretical results in Sections 2 and 3 are derived under the assumption that the covariates are independent. However, the ability of our method to select signal variables from a correlated setting is evidenced by our simulation study and data application. It is interesting to investigate such a setting. Second, we focus in this paper on how this forward method behaves for the CRPS, but the mathematical set-up in Section 2 is much more general and allows to select variables with respect to other loss functions. It would be interesting to extend the current results to a more general set of loss functions.

Acknowledgments

The authors would like to thank Maurice Schmeits, Kiri Whan and Dirk Wolters for many useful discussions on the application of maximum temperature. The authors would like to sincerely thank the three referees and the associate editor for the constructive comments which led to a substantial improvement of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is part of the research project ‘Probabilistic forecasts of extreme weather utilizing advanced methods from extreme value theory’ with project number 14612 which is financed by the Netherlands Organisation for Scientific Research (nederlandse organisatie voor wetenschappelijk onderzoek).

ORCID

Jasper Velthoen http://orcid.org/0000-0003-0419-2508
Juan-Juan Cai http://orcid.org/0000-0002-3788-8125
Geurt Jongbloed http://orcid.org/0000-0003-4708-5868

References

[1] S. Athey, J. Tibshirani, and S. Wager, et al. *Generalized random forests*, Ann. Stat. 47 (2019), pp. 1148–1178.

[2] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32.

[3] A.P. Dawid, *Present position and potential developments: Some personal views: Statistical theory: The prequential approach*, J. R. Stat. Soc. Ser A (General) 147 (1984), pp. 278–292.

[4] E.S. Epstein, *A scoring system for probability forecasts of ranked categories*, J. Appl. Meteorol. (1962–1982). 8 (1969), pp. 985–987.

[5] E.W. Fox, R.A. Hill, S.G. Leibowitz, A.R. Olsen, D.J. Thornbrugh, and M.H. Weber, *Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology*, Environ. Monit. Assess. 189 (2017), pp. 1–20.

[6] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, *Variable selection using random forests*, Pattern. Recognit. Lett. 31 (2010), pp. 2225–2236.

[7] T. Gneiting and A.E. Raftery, *Strictly proper scoring rules, prediction, and estimation*, J. Am. Stat. Assoc. 102 (2007), pp. 359–378.

[8] T. Gneiting, A.E. Raftery, A.H. Westveld III, and T. Goldman, *Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation*, Mon. Weather Rev. 133 (2005), pp. 1098–1118.

[9] B. Gregorutti, B. Michel, and P. Saint-Pierre, *Correlation and variable importance in random forests*, Stat. Comput. 27 (2017), pp. 659–678.

[10] H. Hersbach, *Decomposition of the continuous ranked probability score for ensemble prediction systems*, Weather Forecast. 15 (2000), pp. 559–570.

[11] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, 2003.

[12] S. Kruizinga, *Objective classification of daily 500 mbar patterns*, in *Preprints sixth conference on probability and statistics in atmospheric sciences*, Vol. 9, American Meterological Society Boston, MA, 1979, p. 12.

[13] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, *Understanding variable importances in forests of randomized trees*, Adv. Neural. Inf. Process. Syst. 26 (2013), pp. 431–439.

[14] N. Meinshausen, *Quantile regression forests*, J. Mach. Learn. Res. 7 (2006 Jun), pp. 983–999.

[15] R.E. Morss, J.L. Demuth, and J.K. Lazo, *Communicating uncertainty in weather forecasts: A survey of the us public*, Weather Forecast. 23 (2008), pp. 974–991.

[16] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, *Conditional variable importance for random forests*, BMC Bioinform. 9 (2008), pp. 307.

[17] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, *Bias in random forest variable importance measures: Illustrations, sources and a solution*, BMC Bioinform. 8 (2007), pp. 25.

[18] M. Taillardat, A.-L. Fougères, P. Naveau, and O. Mestre, *Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting*, Weather Forecast. 34 (2019), pp. 617–634.

[19] M. Taillardat, O. Mestre, M. Zamo, and P. Naveau, *Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics*, Mon Weather Rev. 144 (2016), pp. 2375–2393.

[20] S. Wager and S. Athey, *Estimation and inference of heterogeneous treatment effects using random forests*, J. Am. Stat. Assoc. 113 (2018), pp. 1228–1242.

[21] K. Whan and M. Schmeits, *Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods*, Mon Weather Rev. 146 (2018), pp. 3651–3673.

[22] D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Vol. 100, Academic Press, Oxford, 2011.