Dietz, Sebastian J.; Kneringer, Philipp; Mayr, Georg J.; Zeileis, Achim

Working Paper

Low visibility forecasts for different flight planning horizons using tree-based boosting models

Working Papers in Economics and Statistics, No. 2018-11

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

Suggested Citation: Dietz, Sebastian J.; Kneringer, Philipp; Mayr, Georg J.; Zeileis, Achim (2018) : Low visibility forecasts for different flight planning horizons using tree-based boosting models, Working Papers in Economics and Statistics, No. 2018-11, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

This Version is available at:
https://hdl.handle.net/10419/184989

universität innsbruck

Research Platform
Empirical and Experimental Economics

working paper

eeecon
[triple:e:con]

Low visibility forecasts for different flight planning horizons using tree-based boosting models

Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis

Working Papers in Economics and Statistics

2018-11

University of Innsbruck
https://www.uibk.ac.at/eeecon/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 71022
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/eeecon/wopec/

For a list of recent papers see the backpages of this paper.

Low visibility forecasts for different flight planning horizons using tree-based boosting models

Sebastian J. Dietz
Universität Innsbruck

Philipp Kneringer
Universität Innsbruck

Georg J. Mayr
Universität Innsbruck

Achim Zeileis
Universität Innsbruck

Abstract

Low visibility conditions enforce special procedures that reduce the operational flight capacity at airports. Accurate and probabilistic forecasts of these capacity-reducing low-visibility procedure (*lvp*) states help the air traffic management to optimize flight planning and regulation. In this paper we investigate nowcasts, medium-range forecasts, and the predictability limit of the *lvp* states at Vienna Airport. The forecasts are computed with boosting trees, which consist of an ensemble of decision trees grown iteratively on residuals of previous trees. The model predictors are observations at Vienna Airport and output of a high resolution and an ensemble numerical weather prediction (NWP) model. Observations have highest impact for nowcasts up to a lead time of two hours. Afterwards a mix of observations and NWP forecast variables generates the most accurate predictions. With lead times longer than eight hours NWP output dominates until the predictability limit is reached at +12 days. For lead times longer than two days ensemble output generates higher improvement than a single higher resolution. The most important predictors for lead times up to +18 hours are observations of *lvp* and dew point depression, as well as NWP dew point depression. At longer lead times dew point depression and evaporation from the NWP models are most important.

Keywords: aviation meteorology, statistical forecast, visibility, ceiling, decision tree, boosting.

1. Introduction

Low visibility conditions require special procedures to ensure flight safety at airports. These procedures slow down the air traffic and result in a reduction of the operational airport capacity leading to mean economic loss for airports and airlines. In this study we therefore generate predictions of low visibility at thresholds that directly connect to the capacity-reducing procedures at Vienna Airport. Accurate nowcasts of these low visibility thresholds can help to reorganize flight plans and reduce the economic losses. These forecasts, however, are not only important for flight plan reorganizations. They do also have impact on long-term flight planning to avoid expensive short-term reorganizations. This paper therefore focuses on nowcasts with lead times from +1 hour to +18 hours, and on medium-range forecasts up to +14 days lead time. Additionally we are interested in the predictability limit, which is achieved when the improvement of the forecasts over the climatology vanishes.

Generally, low visibility forecasts are generated with two different approaches (Gultepe *et al.* 2007). The first one is physical modeling and uses relevant physical equations to produce predictions in a defined model area. The second approach, statistical modeling, computes relations between the forecast variable and possible predictor variables from past data. Predictions are produced by applying the relationships to new data. An advantage of this approach is low computational cost and the possibility to directly forecast special quantities, such as visibility classes responsible for capacity reductions.

Statistically-based visibility forecasts were investigated first by Bocchieri and Glahn (1972) using a multiple linear regression approach to forecast ceiling continuously and at several thresholds. The predictor variables of their forecasting model were output of a numerical weather prediction (NWP) model. Based on this model approach Vislocky and Fritsch (1997) produced forecasts of multiple binary thresholds of ceiling and visibility. By adding observations to the model predictors they enhanced the performance at short lead times. This forecasting system was improved by Leyton and Fritsch (2003, 2004) by increasing the density and frequency of the surface observations. Ghirardelli and Glahn (2010) used multiple linear regression to generate an operational prediction system for several visibility and ceiling thresholds for multiple locations and lead times. A comparison of various statistical methods to forecast the same information as Ghirardelli and Glahn (2010), however in one combined variable, was conducted by Herman and Schumacher (2016). They compared K-nearest neighbor, gradient boosting, random forest, and support vector machine methods and found that no specific algorithm performs best overall.

The operationally relevant visibility information for flight management is the low-visibility procedure (*lvp*) state, a combination of visibility and ceiling which directly connects to capacity reductions at airports. It was forecasted first by Kneringer *et al.* (2018) and Dietz *et al.* (2018), who used ordered logistic regression and decision tree based models for observation-based nowcasts up to two hours lead time. Their forecasts are most relevant for short-term regulations. In order to conduct flight plan reorganizations the air traffic management requires forecasts with lead times up to +18 hours, and even longer forecasts are required for long-term flight planning.

The focus of this paper is therefore on determining the skill and most important model predictors for *lvp* nowcasts up to a lead time of +18 hours and for medium-range forecasts from one day up to the predictability limit. For forecast generation we use boosting trees based on current observations and NWP model output. The forecasts are produced for Vienna Airport and only between September and March at 6 UTC since both, *lvp* occurrence probability and arrival rate are highest during this time (Kneringer *et al.* 2018; Dietz *et al.* 2018). The paper is organized as follows: Sect. 2 describes the data sources, the response, and the predictor variables used in this study. Afterwards the statistical methods are explained and the results are analyzed and discussed.

2. Data

Six years of data (November 2011 to November 2017) are available to produce and evaluate forecasts, which result in 1177 observations when considering the cold season (October to March) at 6 UTC only. The forecasts are developed for one specific touchdown point at Vienna Airport and consist of observations at Vienna Airport and NWP model output. All

observations used are measured close to the examined touchdown point.

The NWP model data used for forecast generation are from the atmospheric high-resolution model (HRES) and the ensemble prediction system (ENS) of the European Center for Medium-Range Weather Forecasts (ECMWF). The HRES model provides forecasts with hourly output until a lead time of +90 hours. Afterwards the output is three hourly resolved until +144 hours and six hourly up to the maximum lead time of +240 hours. This model is initialized daily at 00 and 12 UTC and provides one forecast for each lead time. Additionally the ECMWF boundary condition program initializes this model daily at 06 and 18 UTC and generates hourly resolved forecasts up to +90 hours lead time. The horizontal model resolution is 0.1° x 0.1° in latitude/longitude direction, which conforms to grid boxes of approximately 9 x 9 km^2. During the training period the model was improved several times (changes in the horizontal and vertical model grid and the data assimilation scheme). A bilinear interpolation from the four closest grid points to the validation point, however reduces the impact of model grid changes.

The ENS provides forecasts until +15 days (+360 hours) lead time with three-hourly output until +144 hours and six hourly output afterwards. Instead of only one forecast with each output, the ENS provides 50 forecasts (members) at each lead time. Each of the members is computed with slightly changed initial conditions resulting in a different prediction. Using the output information of the 50 members separately would make the predictor variable setup of the statistical models overly large. Thus, only information of the mean and standard deviation of the 50 members is used for forecast generation (Wilks and Hamill 2007; Hamill *et al.* 2008). The ENS is initialized daily at 00 and 12 UTC on a global grid with a 0.2° x 0.2° spatial resolution, which conforms to grid boxes of approximately 18 x 18 km^2. Similarly as the HRES the ENS was improved several times during the model training period. The utilization of a bilinear interpolation again reduces the impact of model grid changes due to the output quality.

2.1. Forecast variable

The response is the low-visibility procedure (*lvp*) state, which is an ordered categorical variable that comes into effect when certain horizontal and/or vertical visibility thresholds are crossed at airports. The horizontal visibility thresholds are determined by observations of the runway visual range (*rvr*), defined as the distance over which the pilot of an aircraft on the centerline of the runway can see the runway surface markings or the lights delineating the runway or identifying its center line (International Civil Aviation Organization 2005). The vertical visibility thresholds are determined by ceiling (*cei*) observations. Ceiling is the base altitude of a cloud deck covering at least five octa of the sky.

The number of *lvp* states and their threshold values vary with the location, size, and technical equipment of an airport. Vienna Airport has four different *lvp* states. Tab. 1 states their thresholds, related capacity reductions and climatological occurrences. Since no restrictions (*lvp*0) occur in about 90 % of the cold season (Oct–Mar) and *lvp*2 is four times more frequent than the less restrictive state *lvp*1 and the maximum restrictive state *lvp*3, forecasts are challenging.

Table 1: Definition of the *lvp* states with their thresholds in runway visual range (*rvr*) and ceiling (*cei*), their climatological occurrence probability, and their maximum operational capacity utilization for Vienna Airport. The climatological occurrence probability is computed during the cold seasons (Oct. – Mar.) from November 2011 to November 2017 at 06 UTC.

lvp state	*rvr*		*cei*	occurrence	capacity
0				89.7%	100%
1	< 1200 m	or	< 90 m	1.7%	75%
2	< 600 m	or	< 60 m	7.1%	60%
3	< 350 m			1.5%	40%

2.2. Predictor variables

The model predictors consist of observations and output of NWP simulations. The observations used are the predictors Kneringer *et al.* (2018) found to have the highest impact on nowcasts (see Tab. 2, top). The horizontal visibilities *vis* and *rvr*, which are both used as predictors differ in the inclusion of background luminance and runway light quality, as well as the truncation at 2000 m for *rvr* (Federal Aviation Administration 2006). Ceiling (*cei*) is post-processed from ceilometer outputs (Dietz *et al.* 2018). The *lvp* state is computed by thresholds of *cei* and *rvr* as described in Sect. 2.1. Dew point depression (*dpd*) and temperature difference between 2 m and 5 cm agl. (*dts*) are computed from temperature sensors in close distance. The climatological information used as predictor is the solar zenith angle (*sza*) in order to capture the annual cycle.

The NWP model outputs used as predictors (Tab. 2, bottom) are selected based on physical mechanisms of fog and cloud formulation and the results of Herman and Schumacher (2016). Each variable is internally derived by the ECMWF from the physical model equations using various physical and statistical relationships. Additionally dew point depression (dpd_{model}) and temperature difference between 2 m and surface (dts_{model}) are computed from the NWP model output 2 m temperature, dew point, and surface temperature.

3. Statistical framework

The forecasts in this paper are generated with boosting trees, which Dietz *et al.* (2018) showed to have comparable performance with ordered logistic regression and better performance than single decision trees, bagging trees, and random forests in case of *lvp* state nowcasts. In combination with decision trees boosting is a flexible method to model additive and nonlinear data features. In the following we describe the characteristics and properties of the boosting trees, the validation criterion, and the reference forecasts used to determine the model benefits.

3.1. Boosting trees

To forecast the *lvp* state we require a model that is able to deal with ordered response variables. Convenient models for such purposes are decision trees, which are intuitive and easy to build, descriptively, and able to model interactions between multiple predictors. Single decision trees, however, have a high variance and in many cases a weak forecast performance

Table 2: Observations, climatological information (top), and NWP model output (bottom) used as predictors for the statistical models. The particular predictors from the ENS consist of the mean and standard deviation of all members.

Variable	Unit	Description
lvp	[0,1,2,3]	low-visibility procedure state
rvr	[m]	runway visual range
vis	[m]	visibility
cei	[m]	ceiling
dpd	[°C]	dew point depression 2 m agl.
dts	[°C]	temperature difference 2 m to 5 cm agl.
sza	[°]	solar zenith angle
bld	[Jm^{-2}]	boundary layer dissipation
e	[m.w.e.][a]	evaporation
cdir	[Jm^{-2}]	clear sky direct solar radiation
dts_{model}	[°C]	temperature difference to surface
dpd_{model}	[°C]	dew point depression
blh	[m]	boundary layer height
lcc	[0 – 1]	low cloud cover
shf	[Jm^{-2}]	sensible heat flux
tp	[m]	total precipitation

[a] meter of water equivalent

compared to highly developed statistical models. To overcome this weakness, an ensemble of decision trees can be grown and merged into one model (James *et al.* 2014).

A well-known and powerful approach to build multiple trees and merge them together is boosting, which establishes new decision trees iteratively on information of previous ones. In the first step of a boosting tree algorithm a single decision tree is developed, from which the residual information is computed and a new tree is fitted to it. Afterwards the new tree is merged to the previous one, however with a shrinkage parameter to grow the model slowly and avoid overfitting. In the next step the residual information of the merged model is computed and a new tree is again fitted to it and merged with the shrinkage parameter to the actual model. The process of recomputing the residuals of the model, fitting a tree to it and merging the new tree to the previous ones is repeated until a stopping criterion is reached (e.g., the number of maximum iterations is reached, the forecast accuracy cannot be improved anymore, etc.).

The boosting algorithm used in this study is the component-wise gradient boosting algorithm of Bühlmann and Hothorn (2007), which computes residual information from the the negative gradient vector of the loss-function of the actual model. In case of ordinal response variables, such as *lvp*, the log-likelihood of the proportional odds model of Agresti (2003) is specified as loss-function (Schmid *et al.* 2011). The decision trees used in this boosting algorithm are established with the unbiased recursive partitioning algorithm of Hothorn *et al.* (2006).

For computation we used the R package **mboost** which implements the described algorithm

(Hothorn *et al.* 2017a). The number of trees for each model is determined by the minimized out-of-sample error. Therefore the model score is computed for each iteration up to a maximum number of 5000 iterations. The particular model for the iteration with the minimum score is then selected. The number of iterations differ for the various cross-validated training samples and for different lead times. The decision trees used in the boosting algorithm are implemented in the R package **party** (Hothorn *et al.* 2017b).

3.2. Model verification

The performance of probabilistic forecasts of ordered response variables such as *lvp* can be assessed by the ranked probability score (RPS; Epstein 1969; Murphy 1971; Wilks 2011). For its computation the squared differences between the cumulative probabilities of the forecast and observation are calculated for each category and summarized afterwards.

$$\mathrm{RPS} = \frac{1}{I-1} \sum_{s=1}^{I} \left[\sum_{i=1}^{s} y_i - o_i \right]^2,$$

with the forecast probabilities y_i and observations o_i for each category $i = 1, \ldots, I$. A perfect forecast results in an RPS of 0, the worst possible forecast in an RPS of 1. For model comparison, the ranked probability scores of each forecast observation pair in the test sample are averaged.

The model RPS is computed out-of-sample by a season-wise cross-validation approach with error bootstrapping. The data set is divided into six blocks, each of which contains data from one cold season. Afterwards, the models are fitted on five blocks and tested on the remaining one until each block is used once for model testing.

Bootstrapping is used to assess model uncertainty. We generate 1000 data samples, each with randomly drawn forecast-observation pairs of the original sample with replacement. The size of each sample is identical to the overall number of forecast-observation pairs. After bootstrapping the mean RPS is computed for each sample. The distribution of the mean ranked probability scores describes the model uncertainty.

3.3. Variable importance measurement

Permutation tests are used to determine the predictors with the highest impact on the forecasts. Such tests compute and compare the forecast performance of the original test sample and of multiple modified versions of it. In the beginning of the test the out-of-sample performance of the original test sample is computed. After predictions on the original sample one predictor variable of the original sample is permuted randomly and new predictions, again with the same model, are generated from this modified sample. Since the new predictions are generated with random information of one predictor they have a loss in forecast performance. The stronger the loss in forecast performance the higher is the impact of the permuted predictor. The procedure of permuting the values of one predictor variable and computing the performance of this modified sample is repeated for each predictor.

Moreover, to extract meaningful information of the most important predictors the permutation test is conducted on each cross-validated test sample. Afterwards the results from the different test samples are averaged to show the mean impact of each predictor on the forecast.

3.4. Reference models

The benefit of the forecasts of the statistical models can be assessed by different reference models. A widely used reference for short lead times is the persistence model (Kneringer *et al.* 2018). The persistence forecast is deterministic and assumes that the *lvp* state does not change between forecast initialization and validation.

For long lead times climatology becomes competitive and is thus used as reference. Climatology always predicts the distribution of the response in the training sample.

The third reference is direct output of the ECMWF NWP model. The ECMWF contains *vis* in its output from May 2015 on and *cei* since November 2016. Thus, the predicted *lvp* state can be computed directly from the NWP model output after November 2016. For the HRES model only deterministic *lvp* state forecasts can be computed because the model consists of one member only. The ENS model, however, consists of 50 members and therefore probabilistic forecasts can be derived by merging the predictions of all 50 members. Comparisons with direct NWP output can be conducted for one cold season.

4. Results

4.1. Nowcasts (+1h to +18h)

This section is about *lvp* state forecasts with lead times from +1 to +18 hours. The predictors for the statistical models are observations and output of the ECMWF HRES model – both separately and combined. The performance of each model is compared amongst others and to the references persistence and climatology. Moreover the predictors with highest impact on the forecasts are examined and analyzed on their effects.

Model performance

The performance of the boosting trees with different predictor setups is given in Fig. 1 for the lead times +1 hour to +18 hours. The statistical models based on observations outperform the persistence reference model already at the shortest lead time. During nowcasts up to +2 hours lead time the difference between observation-based models and persistence is smallest and increases with longer lead times. A longer distance between forecast initialization and validation leads to a higher probability of changing *lvp* states and therefore to a worsen of the persistence. Similarly the relations of current observations and future *lvp* decrease with longer lead times and the observation-based models converge to climatology, however much slower than persistence.

The boosting trees based on the HRES output also outperform climatology up to +18 hours lead time. Their performance is constant for the lead times +1 hour to +6 hours because of identical HRES information. In this investigation we assume that NWP model output is available immediately after model initialization. The HRES model is initialized daily at 00, 06, 12, and 18 UTC. The closest output available for the 06 UTC forecast with a lead time of +1 hour is from the 00 UTC initialization with a lead time of +6 hours. This information is used for the lead times +1 to +6 hours. The same applies for the lead times +7 to +12 hours and +13 to +18 hours (with outputs from the 18 and 12 UTC model initialization, respectively). Similarly to the observation-based models, the performance of HRES-based ones decreases

with longer lead times. The decreasing process, however, is much slower. Persistence performs better than HRES-based models only up to +2 hours lead time. Between the lead times +3 and +7 hours the performance of the HRES-based models caught up the observation-based ones. Observation-based models therefore perform on average better until a lead time of +5 hours (Fig. 1b).

The best performing models are the ones with the combined predictor setup. With nowcasts up to +2 hours lead time they perform almost identically to observation-based models. During the lead times +3 to +7 hours they perform better than both other models. Primarily they perform similarly to observation-based models and converge slowly to the performance of the HRES-based models.

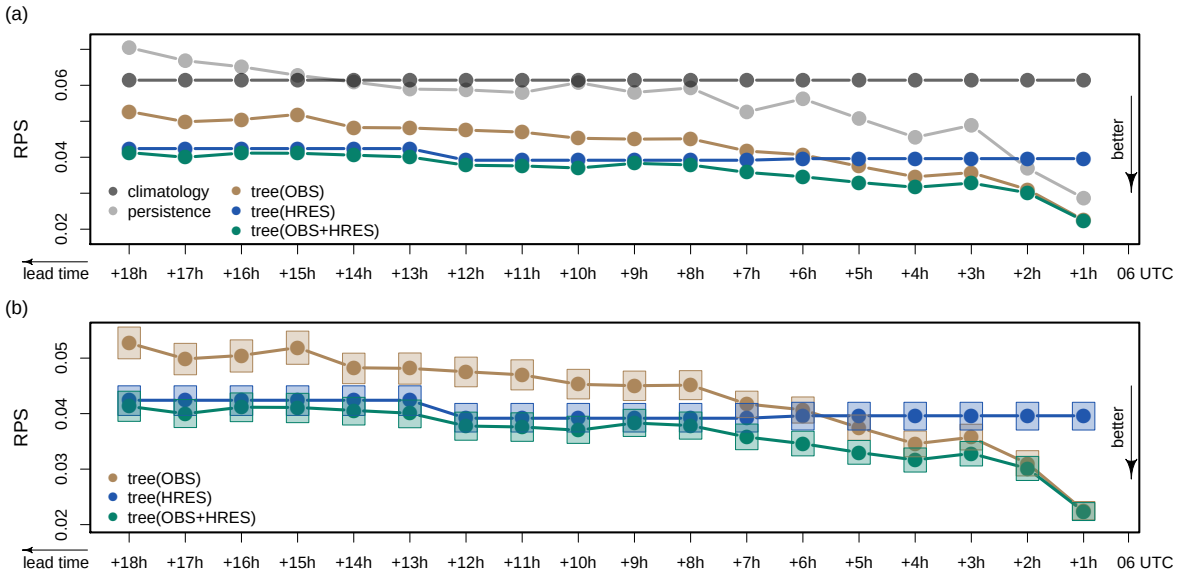

Figure 1: Forecast performance of boosting tree models based on observations (OBS), NWP model output of the deterministic HRES ECMWF model (HRES), and the combined predictor variable setup (OBS+HRES). The forecast validation time is always 6 UTC. Models with a lead time of +1h (+2h, ...) are initialized at 5 UTC (4 UTC, ...). The dots show the mean performance of the particular model. Lower RPS indicates a better forecast performance. (a) Mean performance of the statistical models and the references climatology and persistence. (b) Forecast performance of the statistical models with their uncertainty (25–75 percentile range).

Impact of predictors

The predictors with highest impact on the forecast are analyzed using the variable permutation test (Sect. 3.3) applied to the boosting trees with the combined predictor setup. Figure 2 shows the predictors with highest impact on forecasts for the lead times +1, +6, and +12 hours.

Forecasts with the lead time +1 hour mainly rely on observations. This relation can also be seen in the analysis of the forecast performance in Fig. 1b, where models with the mixed variable setup perform nearly identically to observation-based models. The *lvp* state at forecast

initialization has the highest impact amongst all predictors. If this value would be random the *lvp* performance worsens in average 124%. Observed dew point depression (*dpd*), runway visual range (*rvr*), and visibility (*vis*) have also impact while the solar zenith angle (*sza*) contributes little information. These findings agree with the results of Kneringer *et al.* (2018) and Dietz *et al.* (2018).

The impact of observations decreases strongly for nowcasts with lead times from +3 to +7 hours. Dew point depression from the NWP model (dpd_{model}) and observations of *lvp* have the highest impact at +6 hourly forecasts. Further variables, albeit with smaller impact are observations of dew point depression and visibility, and evaporation (e) from the NWP model output.

As the forecasting horizon increases from +8 to +18 hours the influence of dew point depression from the NWP model increases, whereas other predictors only have small impact. Random *lvp* states at forecast initialization, for example, would decrease the performance by less than 5% for predictions with +12 hours lead time. The performance of the models with the combined predictor setup is similar to the performance of the HRES-based models. The strong influence of NWP model-based dew point depression on the forecast performance confirms this finding.

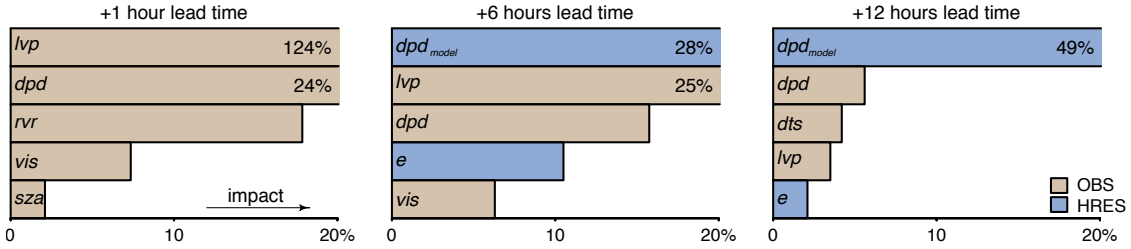

Figure 2: Predictors of Tab. 2 with highest impact on boosting trees with the combined predictor setup for lead times of +1, +6, and +12 hours, respectively. The dark yellow color indicates observation-based predictors, the blue color HRES-based ones. The x-axis (numbers in the bars) shows the percentage decrease in forecast performance when the true values of the particular predictor are replaced with random information.

The effect of particular predictors on the forecast probabilities is shown in Fig. 3. To illustrate the effects of the predictors, they are fixed to certain, preselected values. Afterwards, the value of one predictor is modified to generate predictions. Shown are the effects of observed and NWP model dew point depression (*dpd*, dpd_{model}), *lvp*, and evaporation (e) for the lead times +6 hours and +12 hours, respectively.

The fixed values were selected for conditions that can drift to both, *lvp*0 events and capacity reducing *lvp* events (Tab. 3). For predictions with +6 hours lead time the fixed values result in 54% *lvp*0, 11% *lvp*1, 32% *lvp*2, 3% *lvp*3, for +12 hours lead time in 57% *lvp*0, 9% *lvp*1, 30% *lvp*2, 4% *lvp*3.

Dew point depression (*dpd*) which has highest impact on the forecast has a positive effect due to *lvp*0 states. Increased *dpd* forces an increased probability of *lvp*0. High *dpd* reflects a dry atmosphere where fog formation is unlikely but dissipation likely.

Observed *lvp* has the reverse effect as dew point depression. High *lvp* states (*lvp*1/2/3) at forecast initialization result in a low probability for upcoming *lvp*0. Conditions with *lvp*1/2/3

reflect a moist atmosphere and the main process for drying the atmosphere is solar radiation, which does not appear during night (forecast validation is always at 06 UTC).

Increasing evaporation from the ground has the same effect as increasing *lvp* states. Condensation forms small water drops close to ground and reduces visibility.

Table 3: Fixed values for the computation of the effects of particular predictors (cf. Tab. 2) on the forecast.

Variable	Value	Variable	Value
lvp	0	*blh*	150 m
rvr	2000 m	dpd_{model}	0.3 °C
vis	2500 m	dts_{model}	0.2 °C
cei	500 ft	*cdir*	0 Jm^{-2}
dpd	0.5 °C	*e*	$6{\cdot}10^{-6}$ m.w.e
dts	0.1 °C	*lcc*	0.9
sza	150 °	*shf*	1500 Jm^{-2}
bld	3000 Jm^{-2}	*tp*	0 m

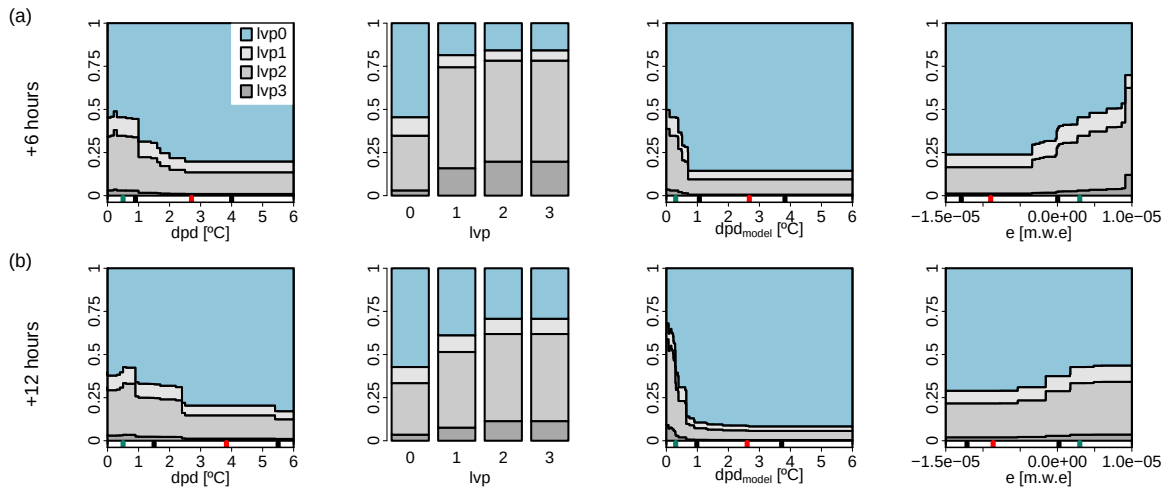

Figure 3: Effects of observed and NWP model dew point depression (*dpd*, dpd_{model}), *lvp*, and evaporation (e) for the lead times +6 hours and +12 hours, respectively. Downward fluxes are defined positive in the ECMWF and therefore positive evaporation represents condensation. The red marks at the x-axis show the median and the black marks the 25 and 75 percentiles distribution of the variable in the data set. The green marks show the fixed values from Tab. 3.

4.2. Medium-range forecasts and predictability limit

The performance of models with the combined predictor setup converges to HRES-based models at lead times longer than +8 hours (Fig. 1). Therefore we only use NWP-model-based predictors for the generation of medium-range forecasts and the investigation of the predictability limit. The predictors used include deterministic information from the HRES and the means and standard deviations from the ENS.

Model performance

Figure 4 shows the performance of boosting trees based on outputs of HRES and ENS for medium-range forecasts with lead times from +0 to +14 days. The predictions consist of output of the 00 UTC NWP model run and the forecast validation time is again 06 UTC. Lead times of +0, +1, +2, etc. days correspond to +6, +30, +54, etc. hours. The maximum output length of the HRES is +240 hours. HRES-based model forecasts can be generated therefore only up to +9 days lead time. The ENS, on the other hand, allows forecasts up to +14 days lead time.

The performance of the statistical models and climatology is shown in Fig. 4a with their uncertainties. In the nowcasting range with a lead time of +0 days HRES-based statistical models perform slightly better than ENS-based ones. From +1 day lead time on both models perform similar and after +2 days lead time the ENS-based models perform better. The biggest difference in forecast performance occurs for the lead times +4 to +6 days, where ENS-based models clearly outperform HRES-based ones, which converge much faster to climatology. The predictability limit, where the forecasts of climatology and the statistical models perform similar in mean RPS, is reached after approximately +12 days lead time.

In order to obtain more information of the benefit of the statistical models we compare them to the raw output of the NWP models. Raw *lvp* state is computed from the NWP model outputs visibility and ceiling. Since ceiling has been only available from November 2016 on, an out-of-sample comparison between the forecasts of the statistical models and the raw NWP model output is computed between December 2016 and November 2017 (cold season only). We therefore train the boosting trees with cold season data from December 2011 to November 2016 and compare their performance with the raw NWP model output for the remaining period.

Figure 4b shows the mean out-of-sample performance of the statistical models, raw NWP model output, and climatology only for cold season data between December 2016 to November 2017. This period had a much higher occurrence of *lvp*1/2/3 than climatologically expected (Fig. 4a).

HRES-based raw output performs better than climatology only until +1 day. Direct output from the ENS, however, has a benefit over climatology until +6 days lead time. The statistical models with input from the ensemble model have a benefit over the raw ENS output up to +14 days lead time and remain better than climatology up to +12 days. Note that all *lvp* cases detected in the individual ensemble members have their origin in low ceiling cases. The ECMWF visibility does not fall below the *lvp* threshold range during the test period. Moreover, raw *lvp* state forecasts from the ensemble average visibility and ceiling always result in *lvp*0. The reason therefore is the exceeding of the *lvp* thresholds in the variable means for the complete data set.

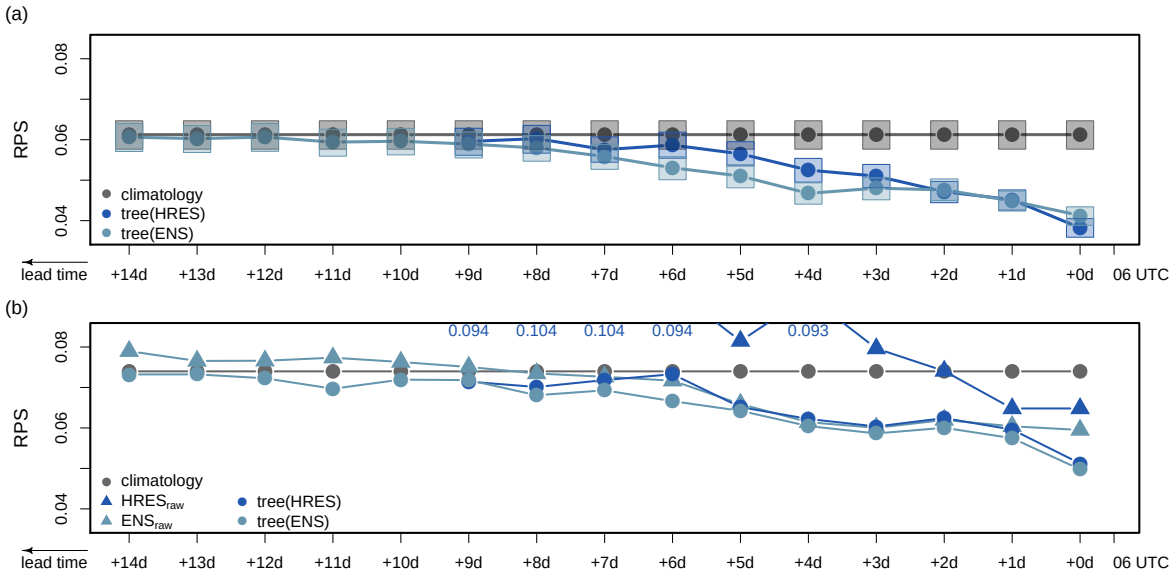

Figure 4: (a) Medium-range forecast performance of the boosting trees based on HRES (tree(HRES)) and ENS (tree(ENS)) information compared to climatology with its uncertainty for the complete 6 years data (cold season only). The model uncertainty is shown with the 25 and 75 percentiles (boxes around the mean). (b) Forecast performance of the statistical models and the reference models climatology and raw NWP model output (HRES_{raw}, ENS_{raw}) only for December 2016 – March 2017 and October – November 2017. The *lvp* state from the raw ensemble is computed by the distribution of the *lvp* states from each member. Computing the *lvp* state only from mean visibility and mean ceiling would always result in *lvp*0. All *lvp* cases from the raw model output are due to low ceiling. The numbers in the upper bounding in Fig. 4b depict the RPS of the raw HRES predictions.

Highest impact inputs

The most important predictors for statistical based medium-range *lvp* forecasts are again analyzed with the variable permutation test. Figure 5 shows the predictors with highest impact for the models based on HRES and ENS for the lead times +2 days and +8 days. In case of the ENS-based models almost only predictors with mean information do have an impact on the forecast. Little information is contained in the standard deviation.

Dew point depression (*dpd*) has highest impact for both models with +2 days lead time. The performance of HRES-based models decreases on average 20% when observations are replaced by random values. Additional impact on the forecast originates from the predictors boundary layer height (*blh*), sensible heat flux (*shf*), evaporation (*e*), and clear sky direct solar radiation (*cdir*).

When the skill of the model forecasts over climatology decreases the number of predictors with impact on the forecast also decreases. In HRES-based models only two predictors do have influence on predictions with +8 days lead times. Moreover, the impact of these two predictors decreases strongly compared to the impacts of the predictors with the +2 days forecast. The convergence of the statistical models to climatology for longer lead times (+8 days in Fig. 4) indicate a low predictability of the predictors used from the NWP models

and therefore no stable association between the NWP output and the upcoming *lvp* state is found by the models. In ENS-based models, which perform better at long lead times, more predictors have influence on the forecasts and the impact of these predictors is generally higher.

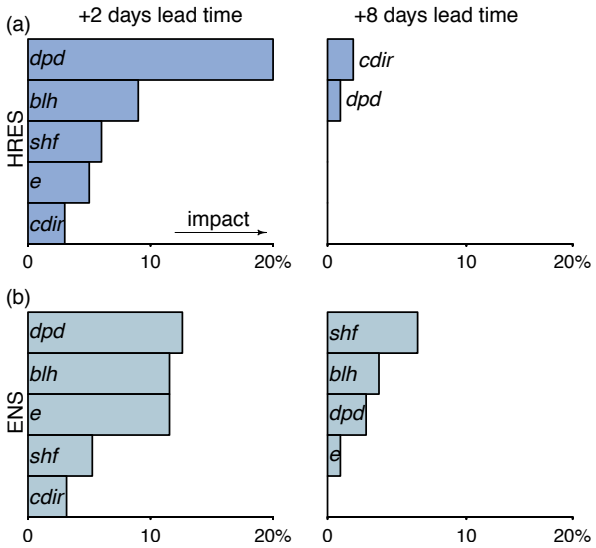

Figure 5: Predictors of Tab. 2 (bottom) with highest impact for medium-range forecasts with +2 days and +8 days lead time. The x-axis shows the percentage decreasing in performance when replacing the true observation of a particular predictor with random information. (a) HRES-based models (b) ENS-based models.

5. Discussion and conclusion

Predictions of *lvp* (low-visibility procedure) states have been developed for flight planning with different horizons using boosting trees. The *lvp* state, which is the relevant variable for flight regularization due to low visibility at airports, is categorical and consists of multiple thresholds of horizontal and vertical visibility. Former studies predict the horizontal and vertical visibility separately, which then can be combined by the air traffic management (e.g. Vislocky and Fritsch 1997; Ghirardelli and Glahn 2010, etc.). This approach, however, makes accurate *probabilistic* forecasts of the *lvp* state impossible because of the interdependence of both visibility variables. Direct forecasts of the *lvp* states, on the other hand, allow probabilistic predictions of the information relevant for aviation. The *lvp* state predictions in this study focus on time scales for short-term regulation, flight plan reorganization, and long-term flight planning.

Short-term regulations are defined with predictions up to the next two hours, which are most important for the flight controllers. These forecasts are the most accurate ones and are mainly driven by latest observations of *lvp*, dew point depression, and visibility.

For reorganizations of flight plans the air traffic management can use the predictions with lead times from +3 to +18 hours. Within this range the impact of observations decreases and NWP model output becomes more important. Highly resolved deterministic NWP output leads to

slightly better performance than ensemble information. For forecasts with lead times of +6 hours the NWP model output dew point depression and the observation *lvp* do have equal impact. Hence observations and NWP output have to be included in the statistical models to generate most accurate predictions. The most important predictors are observations of *lvp*, horizontal visibility, dew point depression, air temperature difference between 2 m and surface, and the NWP model outputs dew point depression and evaporation.

Long-term flight planning requires medium-range forecasts with lead times longer than one day. During this time range the statistical models with postprocessed ensemble information perform most accurately. The NWP outputs with highest benefit for the predictions are dew point depression, evaporation, sensible heat flux, and boundary layer height. The predictability limit of *lvp* is approximately 12 days, where the benefit of the statistical forecasts over climatology vanishes.

The ECMWF NWP models also provide information of visibility and ceiling. Both variables can be used to predict *lvp* directly. However, these variables are not included in the statistical models because their data archive is too short. Comparisons between direct *lvp* state forecasts from the NWP models and the statistical models were made for one cold season and just showed a small difference in the performance between +1 and +5 days lead time. Therefore the statistical models always perform somewhat better. The *lvp* state climatology of the comparison period, however, differs strongly to the climatology of the model training period, which suggests a too short comparison period for valuable statements. Nevertheless, for future investigations of the *lvp* state NWP model output of ceiling and visibility should be included in the statistical models to improve the forecast performance. Of both variables, however, information of each particular member should be taken into account instead of mean ensemble information since mean visibility and/or ceiling always leads to *lvp*-free conditions.

In summary we saw that statistically-based probabilistic *lvp* forecasts do have a benefit over all reference models until a lead time of approximately +12 days. These predictions can be used to improve flight planning at all required forecast horizons.

6. Acknowledgments

This study has been supported by the Austrian Research Promotion agency FFG 843457 and the PhD scholarship of the University of Innsbruck. We want to thank Markus Kerschbaum, Andreas Lanzinger, Martin Steinheimer, and the meteorologists at Vienna Airport for helpful discussions. Moreover, we thank the Austro Control GmbH for providing access to the observation data and the Zentralanstalt für Meteorologie und Geodynamik for providing the ECMWF data.

References

Agresti A (2003). *Categorical Data Analysis*. John Wiley & Sons, Inc. ISBN 978-0-470-46363-5.

Bocchieri JR, Glahn HR (1972). "Use of Model Output Statistics for Predicting Ceiling Height." *Monthly Weather Review*, **100**(12), 869–879. doi:10.1175/1520-0493(1972)100<0869:UOMOSF>2.3.CO;2.

Bühlmann P, Hothorn T (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting." *Statistical Science*, **22**(4), 477–505. doi:10.1214/07-STS242.

Dietz SJ, Kneringer P, Mayr GJ, Zeileis A (2018). "Forecasting Low-Visibility Procedure States with Tree-Based Statistical Methods." *Pure and Applied Geophysics*, **in press.** doi:10.1007/s00024-018-1914-x.

Epstein ES (1969). "A Scoring System for Probability Forecasts of Ranked Categories." *Journal of Applied Meteorology*, **8**, 985–987. doi:10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Federal Aviation Administration (2006). "Performance Specification PC Based Runway Visual Range (RVR) System." *Technical Report FAA-E-2772B*, Department of Transportation. URL https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/techops/navservices/lsg/rvr/media/FAA-E-2772B.pdf.

Ghirardelli JE, Glahn B (2010). "The Meteorological Development Laboratorys Aviation Weather Prediction System." *Weather and Forecasting*, **25**(4), 1027–1051. doi:10.1175/2010WAF2222312.1.

Gultepe I, Tardif R, Michaelides SC, Cermak J, Bott A, Bendix J, Müller MD, Pagowski M, Hansen B, Ellrod G, Jacobs W, Toth G, Cober SG (2007). "Fog Research: A Review of Past Achievements and Future Perspectives." *Pure and Applied Geophysics*, **164**(6), 1121–1159. ISSN 1420-9136. doi:10.1007/s00024-007-0211-x.

Hamill TM, Hagedorn R, Whitaker JS (2008). "Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation." *Monthly Weather Review*, **136**(7), 2620–2632. doi:10.1175/2007MWR2411.1.

Herman GR, Schumacher RS (2016). "Using Reforecasts to Improve Forecasting of Fog and Visibility for Aviation." *Weather and Forecasting*, **31**(2), 467–482. doi:10.1175/WAF-D-15-0108.1.

Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B (2017a). *mboost: Model-Based Boosting*. R package version 2.8-0, URL https://CRAN.R-project.org/package=mboost.

Hothorn T, Hornik K, Strobl C, Zeileis A (2017b). *party – A Laboratory for Recursive Partytioning*. R package version 1.2-3, URL http://CRAN.R-project.org/package=party.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. doi:10.1198/106186006X133933.

International Civil Aviation Organization (2005). "Manual of Runway Visual Range Observing and Reporting Practices." *Technical Report Doc 9328 AN/908*. URL http://dgca.gov.in/intradgca/intra/icaodocs/Doc%209328%20-%20Manual%20Runway%20Visual%20Range%20Observing%20and%20Reporting%20Ed%203%20%20Amd%201%20(En).pdf.

James G, Witten D, Hastie T, Tibshirani R (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics, New York, USA. ISBN 978-1-4614-7137-0.

Kneringer P, Dietz S, Mayr GJ, Zeileis A (2018). "Probabilistic Nowcasting of Low-Visibility Procedure States at Vienna International Airport During Cold Season." *Pure and Applied Geophysics*, **in press.** doi:10.1007/s00024-018-1863-4.

Leyton SM, Fritsch JM (2004). "The Impact of High-Frequency Surface Weather Observations on Short-Term Probabilistic Forecasts of Ceiling and Visibility." *Journal of Applied Meteorology*, **43**, 145–156. doi:10.1175/1520-0450(2004)043<0145:TIOHSW>2.0.CO;2.

Leyton SM, Fritsch M (2003). "Short-Term Probabilistic Forecasts of Ceiling and Visibility Utilizing High-Density Surface Weather Observations." *Weather and Forecasting*, **18**, 891–902. doi:10.1175/1520-0434(2003)018<0891:SPFOCA>2.0.CO;2.

Murphy AH (1971). "A Note on the Ranked Probability Score." *Journal of Applied Meteorology*, **10**, 155–156. doi:10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

Schmid M, Hothorn T, Maloney KO, Weller DE, Potapov S (2011). "Geoadditive Regression Modeling of Stream Biological Condition." *Environmental and Ecological Statistics*, **18**(4), 709–733. doi:10.1007/s10651-010-0158-4.

Vislocky RL, Fritsch MJ (1997). "An Automated, Observations-Based System for Short-Term Prediction of Ceiling and Visibility." *Weather and Forecasting*, **12**, 31–43. doi:10.1175/1520-0434(1997)012<0031:AAOBSF>2.0.CO;2.

Wilks D (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press. ISBN 978-0-1238-5022-5.

Wilks DS, Hamill TM (2007). "Comparison of Ensemble-MOS Methods Using GFS Reforecasts." *Monthly Weather Review*, **135**(6), 2379–2390. doi:10.1175/MWR3402.1.

Affiliation:

Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr
Institute of Atmospheric and Cryospheric Sciences
Faculty of Geo- and Atmospheric Sciences
Universität Innsbruck
Innrain 52
6020 Innsbruck, Austria
E-mail: Sebastian.J.Dietz@gmail.com, Philipp.Kneringer@gmail.com, Georg.Mayr@uibk.ac.at

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstraße 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@uibk.ac.at

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2018-11 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Low visibility forecasts for different flight planning horizons using tree-based boosting models

2018-10 **Michael Pfaffermayr:** Trade creation and trade diversion of regional trade agreements revisited: A constrained panel pseudo-maximum likelihood approach

2018-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model

2018-08 **Lisa Schlosser, Torsten Hothorn, Reto Stauffer, Achim Zeileis:** Distributional regression forests for probabilistic precipitation forecasting in complex terrain

2018-07 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Delegated decision making and social competition in the finance industry

2018-06 **Manuel Gebetsberger, Reto Stauffer, Georg J. Mayr, Achim Zeileis:** Skewed logistic distribution for statistical temperature post-processing in mountainous areas

2018-05 **Reto Stauffer, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Hourly probabilistic snow forecasts over complex terrain: A hybrid ensemble postprocessing approach

2018-04 **Utz Weitzel, Christoph Huber, Florian Lindner, Jürgen Huber, Julia Rose, Michael Kirchler:** Bubbles and financial professionals

2018-03 **Carolin Strobl, Julia Kopf, Raphael Hartmann, Achim Zeileis:** Anchor point selection: An approach for anchoring without anchor items

2018-02 **Michael Greinecker, Christopher Kah:** Pairwise stable matching in large economies

2018-01 **Max Breitenlechner, Johann Scharler:** How does monetary policy influence bank lending? Evidence from the market for banks' wholesale funding

2017-27 **Kenneth Harttgen, Stefan Lang, Johannes Seiler:** Selective mortality and undernutrition in low- and middle-income countries

2017-26 **Jun Honda, Roman Inderst:** Nonlinear incentives and advisor bias

2017-25 **Thorsten Simon, Peter Fabsic, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Probabilistic forecasting of thunderstorms in the Eastern Alps

2017-24 **Florian Lindner:** Choking under pressure of top performers: Evidence from biathlon competitions

2017-23 **Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood

2017-22 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Forecasting low-visibility procedure states with tree-based statistical methods

2017-21 **Philipp Kneringer, Sebastian J. Dietz, Georg J. Mayr, Achim Zeileis:** Probabilistic nowcasting of low-visibility procedure states at Vienna International Airport during cold season

2017-20 **Loukas Balafoutas, Brent J. Davis, Matthias Sutter:** How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior

2017-19 **Martin Geiger, Richard Hule:** The role of correlation in two-asset games: Some experimental evidence

2017-18 **Rudolf Kerschbamer, Daniel Neururer, Alexander Gruber:** Do the altruists lie less?

2017-17 **Meike Köhler, Nikolaus Umlauf, Sonja Greven:** Nonlinear association structures in flexible Bayesian additive joint models

2017-16 **Rudolf Kerschbamer, Daniel Muller:** Social preferences and political attitudes: An online experiment on a large heterogeneous sample

2017-15 **Kenneth Harttgen, Stefan Lang, Judith Santer, Johannes Seiler:** Modeling under-5 mortality through multilevel structured additive regression with varying coefficients for Asia and Sub-Saharan Africa

2017-14 **Christoph Eder, Martin Halla:** Economic origins of cultural norms: The case of animal husbandry and bastardy

2017-13 **Thomas Kneib, Nikolaus Umlauf:** A primer on bayesian distributional regression

2017-12 **Susanne Berger, Nathaniel Graham, Achim Zeileis:** Various versatile variances: An object-oriented implementation of clustered covariances in R

2017-11 **Natalia Danzer, Martin Halla, Nicole Schneeweis, Martina Zweimüller:** Parental leave, (in)formal childcare and long-term child outcomes

2017-10 **Daniel Muller, Sander Renes:** Fairness views and political preferences - Evidence from a large online experiment

2017-09 **Andreas Exenberger:** The logic of inequality extraction: An application to Gini and top incomes data

2017-08 **Sibylle Puntscher, Duc Tran Huy, Janette Walde, Ulrike Tappeiner, Gottfried Tappeiner:** The acceptance of a protected area and the benefits of sustainable tourism: In search of the weak link in their relationship

2017-07 **Helena Fornwagner:** Incentives to lose revisited: The NHL and its tournament incentives

2017-06 **Loukas Balafoutas, Simon Czermak, Marc Eulerich, Helena Fornwagner:** Incentives for dishonesty: An experimental study with internal auditors

2017-05 **Nikolaus Umlauf, Nadja Klein, Achim Zeileis:** BAMLSS: Bayesian additive models for location, scale and shape (and beyond)

2017-04 **Martin Halla, Susanne Pech, Martina Zweimüller:** The effect of statutory sick-pay on workers' labor supply and subsequent health

2017-03 **Franz Buscha, Daniel Müller, Lionel Page:** Can a common currency foster a shared social identity across different nations? The case of the Euro.

2017-02 **Daniel Müller:** The anatomy of distributional preferences with group identity

2017-01 **Wolfgang Frimmel, Martin Halla, Jörg Paetzold:** The intergenerational causal effect of tax evasion: Evidence from the commuter tax allowance in Austria

University of Innsbruck

Working Papers in Economics and Statistics

2018-11

Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis

Low visibility forecasts for different flight planning horizons using tree-based boosting models

Abstract
Low visibility conditions enforce special procedures that reduce the operational flight capacity at airports. Accurate and probabilistic forecasts of these capacity-reducing low-visibility procedure (lvp) states help the air traffic management to optimize flight planning and regulation. In this paper we investigate nowcasts, medium-range forecasts, and the predictability limit of the lvp states at Vienna Airport. The forecasts are computed with boosting trees, which consist of an ensemble of decision trees grown iteratively on residuals of previous trees. The model predictors are observations at Vienna Airport and output of a high resolution and an ensemble numerical weather prediction (NWP) model. Observations have highest impact for nowcasts up to a lead time of two hours. Afterwards a mix of observations and NWP forecast variables generates the most accurate predictions. With lead times longer than eight hours NWP output dominates until the predictability limit is reached at +12 days. For lead times longer than two days ensemble output generates higher improvement than a single higher resolution. The most important predictors for lead times up to +18 hours are observations of lvp and dew point depression, as well as NWP dew point depression. At longer lead times dew point depression and evaporation from the NWP models are most important.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)