Machine-learning regression applied to diagnose horizontal visibility from mesoscale NWP model forecasts

Driss Bari¹ · Abdelali Ouagabi²

Received: 17 January 2020 / Accepted: 25 February 2020 / Published online: 5 March 2020
© Springer Nature Switzerland AG 2020

Abstract

Low-visibility conditions are a weather hazard that affects all forms of transport, and accurate forecasting of their spatial coverage is still a challenge for meteorologists, particularly over a large domain. Current predictions of visibility are based on physical parametrizations in mesoscale models and are thus limited with respect to accuracy. This paper examines the use of supervised machine-learning regression techniques (tree-based ensemble, feed-forward neural network and generalized linear methods) to diagnose visibility from operational mesoscale model forecasts over a large domain. To achieve this, hourly forecasts of meteorological parameters in the lower levels of the atmosphere have been used. In the short-range forecasting framework, the machine-learning algorithms were developed to provide hourly forecasts up to 24 h. To assess the performance of the developed models, hourly observed data, collected at 36 synoptic land stations over the northern part of Morocco, have been used. This region is characterized by a heterogeneous topography. The tree-based ensemble methods have shown some improvement in visibility forecasting in comparison with the operational visibility diagnostic scheme based on Kunkel's formula and also with persistence. It is also found that this machine-learning technique performs better when the forecast depends on multiple predictors instead of only a few with very high importance. In addition, their performance is very sensitive to the disproportionality of data availability between daytime and night-time. Furthermore, it is found that the performance decreases when principal components are used instead of raw correlated data.

Keywords Machine-learning · Visibility · Regression · Principal components · AROME

1 Introduction

Visibility forecasts are important to many customers and are a key requirement for all forms of transport, especially within the fields of land, aviation and shipping, though the range of adverse visibility thresholds of interest varies [1, 2]. Of particular interest is the reliable forecasting of poor visibility, i.e. fog (visibility below 1 km) and mist (visibility between 1 and 5 km) [3]. For example, a dense advection–radiation fog that occurred in December 2013 led to several vehicle collisions and loss of life on a highway of

the Grand Casablanca region, Morocco [4]. For air traffic, a dense fog event (visibility below 200 m) obliged the diversion of 21 aircrafts in 2008, which were supposed to land at Mohamed V international airport, Casablanca, Morocco, to other national airports [5]. Such flight diversion results in a considerable cost to airlines (e.g. additional fuel consumption and passengers support). Thus, high-quality visibility forecasts are important for reducing economic losses and for improving safety by warning drivers/pilots of dangerous conditions that might occur locally anywhere (e.g. road, airport and port) over a large domain. This indicates

✉ Driss Bari, bari.driss@gmail.com | ¹CNRMSI/SMN, Direction de la Météorologie Nationale, 22000 Casablanca, Morocco. ²Research Laboratory in Computer Science and Telecommunications (LaRIT), Department of Computer Science, Faculty of Science, Ibn-Tofail University, Kenitra, Morocco.

the need to improve visibility forecasting, using the available data over this domain, and provide the motivation to carry out this research.

Visibility prediction techniques can be categorized based on numerical weather prediction (NWP), statistical methods as well as integrated methods [1, 6]. Integrated visibility methods are based on the application of conventional rules for estimating the horizontal visibility using the latest atmospheric conditions from a NWP model, together with surface observations for the forecasting location. In this framework, conditional climatology combines persistence and climatology by incorporating knowledge of outcomes of similar weather situations in the past in order to provide short-term forecasts of current weather parameters. However, Vislocky and Fritsch [7] described that persistence had usually better skill for very short range prediction of ceiling and visibility, whereas conditional climatology was a more difficult tool to beat than persistence when the forecast period gets longer [8, 9].

Within the field of numerical weather prediction, visibility is diagnosed as a function of a set of prognostic atmospheric variables, used in a parametrization. The formulation of Kunkel [10] (hereafter K84) is the most common one used in NWP models and is based on the calculation of visibility from liquid water content (e.g. single column model COBEL, [11]; 3D NWP model Meso-NH, [12]). Gultepe et al. [13] formulated an improved parametrization for warm-fog conditions which uses relationships between visibility, liquid water content (LWC) and droplet number concentration. The authors have demonstrated that visibility parametrization based on LWC alone (with a constant droplets concentration N_d) causes a visibility error up to 100% depending on environmental conditions. The Met Office's Unified Model diagnoses the visibility from the prognostic fields of pressure, temperature, specific humidity, cloud liquid water, cloud frozen water and aerosol mass concentration [14]. The author found a root-mean-square error of about 7930 m and bias of -1450 m when he compared neural network-based visibility to the observed one, obtained from the Belfort visiometer at Cardington. All the previous studies indicated that visibility parametrization needs to be improved since it caused an over-/underestimation of visibility depending on environmental conditions.

Statistical supervised machine-learning (ML) techniques rely on a training period of historical data which connect forecasts and observed visibility. In this framework, machine-learning algorithms are emerging as suitable methods for detection and prediction of meteorological phenomena [2, 15, 16]. Some machine-learning methods have been used for low-visibility forecasting, for example artificial networks [17], multiple linear regression [18], and tree-based methods [16]. In the literature, ML regression

techniques have been used for low-visibility events forecasting by Cornejo-Bueno et al. [19] at Valladolid airport, Spain, with a focus on runway visual range. Using ordered logistic regression in the nowcasting framework, Kneringer et al. [20] developed a model to generate probabilistic LVP (low-visibility procedure) categories forecasts at Vienna International Airport. To our knowledge, all the previous machine-learning techniques have been applied to a specific site, particularly the airports. Thus, there are a few studies in the literature regarding the evaluation of the performance of such ML techniques for visibility prediction over a large domain [e.g. [21]]. And, the factors influencing the ML-based model performance need to be more investigated.

This work aims to evaluate the potential of machine-learning regression techniques in diagnosing the horizontal visibility from the forecasts of an operational mesoscale NWP model. The study domain covers the northern part of Morocco (Fig. 1a). This area is characterized by heterogeneous topographical features, dominated by mountainous interiors which reach a peak of 4167 m, crossing diagonally the domain from south-west to north-east and agricultural regions in the south-western part (Fig. 1a). Hourly observational data of horizontal visibility, which are extracted from SYNOP messages issued from 36 synoptic stations, have been used (Fig. 1a) over a 2-year period (from March 2015 to February 2017). These data are used as known values for the target variable in the used supervised learning algorithm. During night-time, 23 synoptic stations, mainly located in airports, are operating during the whole night-time, while the remaining stations operate for a duration up to 5 h during the night (Fig. 1b). Various quality control tests (coherence with the other standard meteorological parameters, missing, etc.) have been applied to the observed visibility, and 87% were classified as correct. Within the total of 326 690 observed data, 64% occurred during the daytime.

Taking into account that horizontal visibility could be also reduced under precipitation [22], we have investigated the frequency of reduced visibility classes that impact human activities. Thus, Fig. 2 illustrates the seasonal (Fig. 2a) and diurnal (Fig. 2b) distributions of reduced visibility below 5 km by fog and mist phenomena, and by precipitation. Figure 2 points out that low visibilities are mainly caused by the occurrence of fog and mist over the 2-year period. It is also seen clearly that reduced visibility below 1 km under rainy conditions is almost never reported over the study domain. The monthly distribution points out the high frequency of mist during the summer season, and the fog frequency reached its peak during the cold season (from October to April). On the other hand, the diurnal distribution shows that the night-time and early morning are the favourable periods for fog occurrence

Fig. 1 **a** Map showing the location of the study domain which covers the northern part of Morocco. **b** Map showing Orography (m) of the study domain and the locations of the synoptic stations. Triangles

refer to stations operating during the whole day, while circles refer to stations which are operating during the daytime and up to 5 h during the night-time

Fig. 2 **a** Seasonal and **b** diurnal distributions of reduced visibilities below 5 km

while the mist can occur at each moment during the daytime. It should be noted that visibilities below 5 km associated with precipitation are often observed during the afternoon although it is of low frequency. From the geographical standpoint, it is found that reduced visibility below 5 km affected rarely more than 4 synoptic stations at the same moment during the day. In this framework, regional-scale visibility forecasts from the current

operational NWP models are limited with respect to accuracy and need to be improved. Can, for example, machine-learning techniques be an alternative to diagnose the horizontal visibility from the NWP model forecasts over a large domain? It is the scope of this study to answer this question based on machine-learning regression algorithms.

Thus, this study aims to generate forecasts of horizontal visibility in the short-range forecasting framework (up

to 24 h) by developing an ML model using the state-of-the-art machine-learning regression techniques. The challenges of this objective include the rare occurrence of reduced visibilities below 5 km over a heterogeneous topographical domain with disproportionality of data availability between daytime and night-time. The quality of the new visibility forecasts will be assessed using a set of independent observations.

Section 2, devoted to data sets and methodology, describes the predictor variables and the 3D NWP mesoscale AROME model (Applications de la Recherche à l'Opérationnel à Méso-Echelle, [23]) used in this study. Machine-learning techniques are also briefly described in this section. In Sect. 3, we analyse the modelling results. A discussion of the main findings with previous studies is conducted in Sect. 4. Conclusions and perspectives are presented in the last section.

2 Predictor variables and methods

2.1 Predictor variables

The mesoscale limited area model used in this study is the AROME model [23], which has been in operational use at the Moroccan National Meteorological Service since 2015. This model is designed to forecast high-impact meteorological phenomena with 2.5 km for the horizontal grid and 60 vertical levels, with the lowest level at about 10 m above the ground. AROME uses the one-moment microphysical scheme ICE3 [24], which is a microphysical parametrization with a fixed cloud droplets concentration ($N_c = 300 \text{ cm}^{-3}$ over the land and $N_c = 100 \text{ cm}^{-3}$ over the sea). Mixing ratios of six types of hydrometeors (cloud, rain, snow, graupel, water vapour, and ice) are included as prognostic variables. This model uses the ECMWF radiation parametrizations [25], and the representation of the turbulence in the planetary boundary layer is based

on a prognostic turbulent kinetic energy equation [26] combined with a diagnostic mixing length. The coupling between the atmosphere and the underlying surface was based on the SURFEX system (<http://www.cnrm.meteo.fr/surfex/>).

In operation, the model is run two times per day, and in this study, the 0000 UTC hourly forecasts (up to 24 h) of every day are used. Table 1 summarizes the list of the AROME parameters used in this study. AROME parameters were selected based on a priori assumptions about parameters that might have the most influence on horizontal visibility at 2 m. To take into account the spatial and temporal distribution of the horizontal visibility over the study domain, the predictor variables include also the geographical coordinates (*Latitude*, *Longitude*, and *Altitude*) of the nearest grid point to each synoptic station and also the month and hour. It should be noted that some uncertainties exist in *Altitude* due the smoothing of model orography in comparison with actual station altitudes. Thus, the final database contains 34 predictors.

2.2 Machine-learning techniques

In general, machine-learning is dedicated to developing and using algorithms that learn from raw data in order to make predictions [27]. These techniques are able to learn nonlinear relationships in data, which give them more power than linear methods. As opposed to the traditional statistics-based methods, machine-learning techniques are data-driven self-adaptive methods, in which there are few a priori assumptions about the models for problems under study. In the literature, there is a variety of data-mining algorithms, which can be classified into two major classes: supervised methods like decision trees and neural network, and unsupervised methods like clustering and associations. In this study, the following five supervised learning algorithms are used: feed-forward neural network algorithms (FFN), generalized

Table 1 AROME atmospheric variables used in the data-mining process

10-m pressure	10-m wind speed	Mean sea level pressure
10-m zonal wind	2-m temperature	Total cloudiness
10-m meridional wind	2-m relative humidity	Convective cloudiness
10-m temperature	2-m solar radiation	High-level cloudiness
10-m liquid water content	2-m thermal radiation	Middle-level cloudiness
10-m cloud fraction	2-m latent heat flux	Low-level cloudiness
10-m kinetic turbulent energy	2-m sublimation flux	Month
10-m relative humidity	2-m sensible heat flux	Hour
10-m wet potential temperature	Surface pressure	Latitude
10-m dry potential temperature	Downward solar radiation	Longitude
10-m vertical velocity	Downward thermal radiation	Altitude
10-m wind direction		

linear regression algorithms (GLM), and the following tree-based ensemble algorithms: gradient boosting machine (GBM), random forest (RF) and extreme gradient boosting (XGB).

Generally, machine-learning uses an approach that assumes a model defined up to a set of parameters, i.e. $y = f(x|\theta)$ where $f(\cdot)$ is the model and θ refers to its parameters, x and y are, respectively, the input and the output of the model. Then, the machine-learning program optimizes the parameters, θ , such that the approximation error is minimized, that is, our estimates are close as possible to the correct values given in the training set. In the supervised learning problems, we try to minimize a regularized objective function:

$$f(\theta) = L(\theta) + \Omega(\theta),$$

where training loss term $L(\theta)$ measures how well the model fits on training data and the additional regularization term $\Omega(\theta)$ measures complexity of the model. This approach aims to have predictive and simple models. In the tree-based ensemble methods, data-mining process is learning functions (trees) instead of learning numerical weights (in feed-forward neural network and generalized linear models).

In machine-learning, regression refers to the problem of learning the relationship between some input variables and a quantitative output variable that is iteratively refined using a measure of error in the predictions made by the model. There are a number of techniques for using regression. Generalized Linear Model (GLM) estimates regression models for outcomes following the exponential distribution (Gaussian or gamma) via a link function (identity, logarithm or inverse) [28].

Feed-forward neural network (FFN) is trained with gradient descent using back-propagation to map the relations between inputs and the outcome [29, 30]. In fact, using multiple hidden layers allows a more sophisticated build-up from simple elements to more complex ones and can fit the data more accurately with fewer parameters.

Tree-based ensemble approaches [31] such as boosting and random forest are based on growing an ensemble of decision trees and aggregating them into one model, which typically reduces the variance and improves the forecast accuracy [32]. Being sequential, boosting is based on weak learners and reduces error mainly by reducing bias and also to some extent variance, by aggregating the output from many models. Both extreme gradient boosting (XGB [33]) and gradient boosting machine (GBM [34]) follow the principle of gradient boosting. There are, however, some differences in modelling details. Specifically, XGB used a more

regularized model formalization to control over-fitting, which gives it better performance. On the other hand, random forest (RF [35]) grows trees in parallel (low bias, high variance). It tackles the error reduction task in the opposite way: by reducing variance. The trees are made uncorrelated to maximize the decrease in variance.

2.3 Experimental set-up

Investigation in the data set points out the existence of imbalance between the occurrence of reduced and good visibilities. Such inequality can cause problems during the development process of ML-based models [36]. Thus, in this work, the sampling of the whole data set into training (70% of all data) and testing (30% of all data) sets has been performed by random split taking into account this condition. This is done to overcome the fact that the developed algorithms may converge on a solution by which the largest class is always forecast [15, 37]. In fact, the data for each location were individually split into subsets of training and testing ensuring that all months, hours, reduced and good visibilities were well represented at 70% in training and at 30% in testing data sets. It should be noted that the days are chosen randomly, and a part of the training set is used during the cross-validation process (fivefold) and the testing set is used as independent data set for model performance assessment. This configuration aims to evaluate the systematic error of the developed algorithms. In other words, we try to measure the ability of each algorithm to benefit from all the data of all the locations to estimate the horizontal visibility over the large domain.

To develop the FFN-based model, two major steps must be taken into account: data preparation for a split into training and testing sets and also the choice of an appropriate architecture of the neural network [38]. Given that the input variables belong to different ranges (e.g. pressure, temperature, relative humidity, etc.), all inputs were scaled to the same range, using a normalization approach, to prevent the FFN algorithm from placing unwarranted emphasis on those inputs with a larger mean magnitude. Regarding the architecture of the neural network, the developed model has 34 inputs and one output, the horizontal visibility. A variety of networks have been tried with one and two hidden layers assuming the rule of Hecht-Nielsen [39] which suggests that maximum of hidden neurons must be $(2n + 1)$ where n is the number of input nodes. After exploring some possible architectures, the configuration 34–39–30–1 with two hidden layers was used in this study. This means that the FFN network is composed of two layers of hidden nodes, one with 39 nodes and the other with 30, with each node in a layer being connected to every node in the next layer.

For the tree-based ensemble techniques, there is no direct practical rule which indicates the set of parameters one should tune and what are the ideal values of these parameters to obtain the optimal output. In this study, a random grid search with cross-validation has been used (H2o platform using R packages, <http://docs.h2o.ai/>).

3 Results

3.1 Models global performance comparison

To evaluate the potential of machine-learning algorithms for estimating the horizontal visibility, the results of the machine-learning experiments on testing data set are presented below, where the estimated value $\hat{f}_i$ is compared to the observed one o_i for N examples of the testing data set. The different developed algorithms are compared using the verification scores listed in Table 2.

Table 3 summarizes the global performance statistics of visibility estimation on the testing data set, issued, respectively, from models based on machine-learning techniques (linear models -GLM-, Tree-based models -GBM, XGB, RF- and neural network-based models -FFN-), and also visibility diagnosed based on physical parametrization (VIS.K84) and persistence (PERS). In PERS model, it is supposed that the visibility expected today for a given hour is the same as that occurred yesterday at the same time. On the other hand, in VIS.K84, horizontal visibility is diagnosed using only liquid water content (LWC) at 2m following Kunkel [10]: $\text{Vis} = a \times \text{LWC}^{-b}$ where a is 0.027 and b is 0.88. The units of LWC and Vis are, respectively, g m^{-3} and km.

Table 3 points out that the generalized linear model is not suitable for visibility estimation over a large domain due to the complexity of such meteorological conditions and the complex interaction between the physical processes. For the other ML techniques, the tree-based methods outperform the FFN method, the persistence, and the Kunkel's based algorithm. The latter has the worst verification scores. The best performance is related to XGB with a

Table 3 Performance statistics of visibility estimation on the testing data set issued from models based on machine-learning techniques (GLM, GBM, XGB, RF, and FFN), and visibility diagnosed based on physical parametrization (VIS.K84)

Algorithm	BIAS (m)	RMSE (m)	MAE (m)	MRE (%)	CC
GLM	- 203	3431	2398	152	0.60
GBM	- 166	2090	1376	49	0.87
XGB	- 144	2013	1322	47	0.88
RF	- 179	2177	1443	54	0.86
DL	- 182	2298	1576	63	0.84
VIS.K84	2933	39,730	6079	155	- 0.02
PERS	125	2886	1493	251	0.77

Persistence (PERS) is also included for comparison

bias of - 144 m, a mean absolute error of 1322 m with 0.88 correlation and a root-mean-square error of 2013 m. The mean absolute relative error is about 47%. Zhang et al. [40] stated that by fusing multimodal information (combining XGBoost and LightGBM), the visibility prediction system can be significantly improved compared to XGBoost- and LightGBM-based models separately. But, the developed model has a higher root-mean-square error (6710 m) in comparison with our finding (2013 m). This could be due to the fact that the authors used only the top ten factors, as model inputs, that are more correlated to visibility among the 38 initial factors. In the following subsection, the machine-learning models will be analysed based on variable importance to identify which predictor variables contribute more during the machine-learning process.

3.2 Model analysis based on variable importance

To identify the variables with the highest impact on the forecasts, the best attributes are selected according to the information gain achieved when the value of that attribute is known. Table 4 shows the most highly represented node variables in the set of trees created during the development process for the different tree-based algorithms. It is clearly from this table that the first five important

Table 2 List of verification scores used in this study to assess the performance of the developed models on the testing data set

verification score	Formulation	Details
Mean bias	$\text{BIAS} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)$	It measures reliability. It is simply the average of the estimated value minus the average observed value
Correlation coefficient	$\text{CC} = \frac{\sum_{i=1}^N (f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^N (f_i - \bar{f})^2 \sum_{i=1}^N (o_i - \bar{o})^2}}$	It is used to measure the relationship of the algorithm output with observation
Mean absolute error	$\text{MAE} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)$	It is the average difference between the algorithm output and the observation
Root-mean-square error	$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2}$	It is used to measure accuracy. It is affected more by larger errors than MAE
Mean relative error	$\text{MRE} = 100 \times \frac{1}{N} \sum_{i=1}^N \frac{ f_i - o_i }{o_i}$	It is used to quantify the fitness of the algorithms used in this research work

Table 4 Variable importance (%) for each ML-developed model

Algorithm	Variable	Percentage of importance (%)	Total (%)
GBM	Longitude	17.79	68.4
	Altitude	17.22	
	10-m pressure	12.17	
	Surface pressure	10.97	
	Latitude	10.25	
XGB	Downward thermal radiation	05.60	24.86
	Mean sea level pressure	05.27	
	10-m pressure	04.70	
	2-m latent heat flux	04.65	
	2-m thermal radiation	04.64	
RF	Altitude	38.03	79.14
	Longitude	15.33	
	Surface pressure	10.20	
	Latitude	08.86	
	2-m relative humidity	06.72	
FFN	10-m potential temperature	09.57	37.72
	Altitude	07.41	
	10-m temperature	07.28	
	Latitude	07.12	
	2-m sublimation flux	06.34	

The five first important variables are shown with their total of importance

variables represent 79.14%, 68.4%, 37.72% and 24.86% of information gain, respectively, for RF, GBM, FFN and XGB. We notice that the geographical positions (*Longitude*, *Latitude*, and *Altitude*) are the highly important variables for RF, GBM, and FFN in association with the pressure at the lowest model level in the atmospheric boundary layer (*Surface pressure*, *10-m Pressure*). It should be noted that variable selection could be a difficult task and more challenging in the presence of highly correlated predictors such as atmospheric fields. This is in agreement with the finding of Dueben and Bauer [41], dealing with challenges for global weather and climate models based on machine-learning. The authors stated that physical knowledge of the properties of the fields, used as input for neural network, and their connectivity are important to design the optimal network architecture. In fact, atmospheric fields such as surface pressure depend on local features of the study domain such as topography. On the other side, for

the best ML-developed model related to XGB method, the first five variables represent only 24.86% with similar relative importance for each parameter. To reach 80% of the total variable importance, XGB used 28 variables in contrary to RF where *Altitude* has a relative importance of 38.03%. This shows that machine-learning techniques perform better when the forecast depends on multiple predictors instead of only a few with very high importance. Based on variable importance, the discrepancy in model analysis between the used ML techniques could be due to the correlation between predictor variables. To check this hypothesis, principal components regression will be performed, in the following subsection, using the same machine-learning techniques.

3.3 Principal components regression

Principal components regression is carried out on a set of possibly collinear features and performs a transformation to produce a new set of uncorrelated features. In this study, the correlation matrix is preferred because we are using time-series data with different measurement units. Thus, the correlation matrix standardizes the used variables [42, 43]. It is considered as the most efficient representation of variance in the data set [e.g. 44]. The number of principal components (PCs) to retain was determined based on variance explained criteria. Thus, one can keep component to account for 90% or to a lesser extent for 80% of the variance [45].

In comparison with the baseline experiment, where the raw correlated data are used, Table 5 summarizes the loss or gain in performance statistics of visibility estimation on the testing data set using 14, 17 and 34 principal components as input predictors and which represent 90%, 95% and 100% of explained variance. Overall, machine-learning techniques applied to principal components show lower performance than when raw data are used as predictors. It is seen clearly from Table 5 that the explained variance rate decreases, as the performance of the developed algorithms deteriorates in terms of bias, root-mean-square error and mean absolute error. For mean absolute relative error and linear correlation, the impact of using principal components as predictors is neutral to negative. On the other hand, using the principal component regression, the XGB method remains the best performing method compared to other techniques. To study in detail spatial and temporal distribution of the error issued from the developed models for estimating visibility over the study domain, the overall performance analysis is complemented in the following subsections by performance analysis separately as a function of forecast hours and location.

Table 5 Performance loss or gain on the testing data set, as difference between the verification score issued from models based on machine-learning techniques (GLM, GBM, XGB, RF, and FFN) using 34, 17 and 14 principal components (PCs) as inputs and that computed for the baseline experiment (REF, in Bold)

Algorithm	PCs	BIAS	RMSE	MAE	MRE	CC
GLM	REF	− 203	3431	2398	152	0.60
GLM	34PC	− 35	809	159	− 4	− 0.11
GLM	17PC	− 35	809	159	− 4	− 0.11
GLM	14PC	− 35	810	159	− 4	− 0.26
GBM	REF	− 166	2090	1376	49	0.87
GBM	34PC	− 18	134	137	− 3	− 0.01
GBM	17PC	− 31	338	297	1	− 0.04
GBM	14PC	− 7	866	659	17	− 0.15
XGB	REF	− 144	2013	1322	47	0.88
XGB	34PC	− 26	191	154	1	− 0.02
XGB	17PC	− 33	323	266	7	− 0.04
XGB	14PC	− 23	881	649	17	− 0.14
RF	REF	− 179	2177	1443	54	0.86
RF	34PC	− 33	226	229	− 2	− 0.03
RF	17PC	− 34	505	411	4	− 0.08
RF	14PC	− 4	931	711	21	− 0.17
FFN	REF	− 182	2298	1576	63	0.84
FFN	34PC	− 29	− 33	− 26	− 5	0.00
FFN	17PC	− 206	185	149	− 8	− 0.02
FFN	14PC	− 254	1012	776	23	− 0.20

Fig. 3 Plot of the performance statistics of visibility estimation on the testing data set issued from models based on machine-learning techniques and persistence, as a function of forecast hour: **a** mean absolute error (MAE, m) and **b** mean absolute relative error (MRE, %)

3.4 Performance as function of forecast hours

Figure 3 represents the evolution of the performance statistics of the different estimators as a function of the forecast hours. This plot shows that the performances of the tree-based algorithms (XGB, GBM, and RF) and the

neural network-based models (FFN) have similar behaviour except for the generalized linear model (GLM) which confirms the worst performance for all forecast hours.

Some discrepancies are noticed in the times series of some verification scores (MRE and MAE) for some ML algorithms (Fig. 3). This could be due to the disproportionality

of the number of synoptic stations operating during the whole day and that do not operate during the night. This could impact, in turn, the training process. To check this hypothesis, we have plotted in Fig. 4 the distribution of MAE and MRE of the different estimators as boxplot for all synoptic stations. The nonparametric Mann–Whitney test was used to explore the null hypothesis that groups of data have the same median at 95% confidence level ([46]). The test was performed on various combinations of groups and has shown a statistically significant difference between the medians for both parameters (MAE and MRE) during the day- and night-time.

By comparing the two verification scores (MAE and MRE) for the different algorithms according to the forecast time during the day (daytime or night-time), Fig. 4 reveals that their behaviours are different. In fact, the median of the MAE of the developed algorithms during daytime is often lower than that found during night-time, with a large dispersion around the median. This could be explained by the scarcity of reduced visibilities below 5 km and the prevalence of good visibility during the daytime, as demonstrated in the climatology study in introduction section. Thus, machine-learning algorithms manage to capture well this dominant class of visibility. This is confirmed by the distribution of the MRE in Fig. 4b where it is found that the MRE is lower during daytime unlike during night-time. This could be due to the uncertainties included in the estimated horizontal visibility, particularly during night-time, by the observer at the synoptic stations. Indeed, the observed visibility is mainly subjective and varies from

observer to observer and then from station to station. Thus, in the following subsection, the performance of the developed machine-learning algorithms as a function of each synoptic station will be investigated.

3.5 Performance as function of spatial location

To evaluate the generalization error of the developed model over the study domain, the verification scores of the unified machine learning-based model, applied to an individual station, have been investigated on the testing set. To establish whether significant differences exist between the different algorithms, some statistics are illustrated in Fig. 5 as box-and-whisker plots for GLM, GBM, XGB, RF, FFN and PERS. The comparison has focused on linear correlation and the mean absolute error between the observed and predicted visibility. The Mann–Whitney test was performed on various combinations of groups and has shown a statistically significant difference between the medians for linear correlation, and no significant difference between the medians for the RMSE, except for GLM algorithm and PERS.

This figure shows that tree-based algorithms (GBM, XGB, and RF) show a positive and strong linear correlation (greater than 0.5 and reached 0.9 for some stations) on more than half of the locations used in this study and that it reaches up to 75% of stations for XGB. Results pointed out that the persistence shows a positive, but weak correlation and even a decorrelation for some stations. On the other hand, the machine-learning algorithms have an

Fig. 4 Box plots of the distribution of **a** mean absolute error and **b** mean absolute relative error of the developed algorithms at day-time and night-time hours on the testing set and persistence as a benchmark. The lower and upper edges of each box refer to the

25th and 75th percentiles, respectively. The horizontal line within each box refers to the median. The lower and upper whiskers are the minimum and maximum values, respectively

Fig. 5 Box plots of the distribution of **a** linear correlation and **b** root-mean-square error of the developed algorithms at the 36 locations on the testing set

RMSE, which varies from 2100 to 3000 m as value for the 75th percentiles, which represents 75% of the synoptic stations, except for GLM, in which this error is much higher. It is concluded that XGB outperforms the other machine-learning algorithms and persistence as a benchmark with high correlation and lowest error over the synoptic stations used in this work.

3.6 Binary scores

In SYNOP weather reports, visibility represents the shortest horizontal visible distance considering all directions (World Meteorological Organization [47]). Visibility is estimated by human observers using landmarks around weather stations. Then, the horizontal visibility is typically classified, but the employed classes vary from place to place. In this study, it is found that the performance of the ML algorithms is affected in low-visibility conditions.

Indeed, the RMSE is better for good visibilities (above 5 km) than for the poor ones (e.g. 1978 m vs. 3272 m for XGB and 2255 m vs. 3820 m for FFN) associated with high linear correlation (e.g. 0.88 vs. 0.45 for XGB and 0.84 vs. 0.42 for FFN).

An alternative method to assess the utility of the ML regression techniques is to compute the scores for ML model's diagnosed visibility either being below 5 km (poor visibility) or not. This gives the test its binary nature, and the results of this are shown in Table 6. Frequency bias (FBIAS) is used to measure the reliability of the developed models while per cent correct (PC), probability of detection (POD), critical success index (CSI) and false-alarm ratio (FAR) are used to measure their accuracy. The equitable threat score (ETS) and true skill score (TSS) are used as skill scores (for more details about the verification scores, see WMO/TD No. 1023).

Table 6 Binary scores for the ML-developed model over the testing data set

Algorithm	FBIAS	POD	FAR	PC	CSI	ETS	POFD	TSS
XGB	1.54	0.55	0.65	0.97	0.27	0.26	0.02	0.53
GBM	1.23	0.45	0.64	0.97	0.25	0.24	0.02	0.43
GLM	0.38	0.11	0.71	0.98	0.09	0.08	0.01	0.1
RF	1.3	0.39	0.7	0.97	0.21	0.2	0.02	0.37
FFN	1.96	0.5	0.75	0.96	0.2	0.19	0.03	0.47
PERS	0.54	0.23	0.58	0.96	0.17	0.16	0.01	0.22
VIS.K84	2.06	0.33	0.84	0.96	0.12	0.11	0.03	0.3

Visibility diagnosed based on physical parametrization (VIS.K84) and persistence (PERS) are also included for comparison

From Table 6, it can be seen that XGB model outperforms the other estimators where it correctly diagnoses 97% of both visibility classes and 55% of visibilities below 5 km. XGB model fails in 65% of poor visibilities in the testing set. This could be due to its frequency bias of 1.54 which indicates a slight over-forecasting of poor visibilities. This ML model has the highest values of CSI (0.27) and ETS (0.26); this indicates that 26% of observed poor visibilities were correctly predicted, adjusted for hits that can occur purely due to random chance. In addition, this model shows an important ability to discriminate between good and poor visibilities (TSS = 0.55) followed by FFN model (TSS = 0.47). The latter over-forecasts almost twice the poor visibilities (FBIAS = 1.97) which induces a high false alarm rate FAR of 75%.

4 Discussion

In the short-range forecasting framework (up to 24 h), the suitability of ML approaches for regional-scale visibility forecasts has been investigated in this study. The best performance is related to XGB (extreme gradient boosting) with a bias of -144 m, a mean absolute error of 1322 m with 0.88 correlation and a root-mean-square error of 2013 m. Its mean absolute relative error is about 47%. This result confirms the findings in Caruana and Niculescu-Mizil [48] where tree-based algorithms perform exceptionally well comparing empirically to nine other learning methods in a classification framework. Claxton [14] found an RMSE of about 7930 m and bias of -1450 m when he compared neural network-based visibility to the observed one, obtained from the Belfort visiometer at Cardington. On the other side, results demonstrate that visibility estimation using Kunkel's parametrization has the worst performance with a mean relative error of 155%. This finding is in agreement with that of Gultepe et al [13] where the authors have demonstrated that visibility parametrization based on LWC alone (with a constant droplets concentration N_d) causes a visibility error up to 100% depending on environmental conditions [49].

Regarding the potential of machine-learning techniques to be used as an alternative for diagnosing the horizontal visibility from mesoscale model forecasts, it is found that XGB (extreme gradient boosting) is a suitable tool for estimating visibility over a large domain. Investigation of the performance of the ML-developed models, applied to an individual station, demonstrates that XGB outperforms the other machine-learning algorithms and persistence as a benchmark with high correlation and lowest error over the used synoptic network. In fact, Hsieh and Tang [50] stated, in their review article, that neural networks have difficulty being adapted to analyse or forecast

some meteorological phenomena when using large spatial data fields which cover numerous spatial grids or stations. Thus, neural networks will have a very large number of weights (a combination of the number of inputs and the number of hidden neurons) to deal with during the optimization process. This could lead to an ill-conditioned problem. To overcome this problem, the authors suggest using principal components as a prefilter for compressing the large spatial data fields. To check this in our framework, the robustness of the algorithms to the features correlation has been investigated, and it is found that the performance of ML techniques decreases when principal components are used instead of raw correlated data.

One strength of the tree-based methods is the interpretability of the nonlinear relations through peering inside the black box via the investigation of features importance. In this framework, it is demonstrated here that ML-developed algorithms perform better when the forecast depends on multiple predictors instead of only a few with very high importance. Indeed, models analysis based on variable importance shows that XGB uses 28 predictor variables to reach 80% of importance during the data-mining process, while it is found that changing meteorological conditions over the large domain affected visibility forecasts for the other ML techniques. Indeed, it is found that geographical positions (*Longitude*, *Latitude*, and *Altitude*) are the variables with the highest impact on the forecasts for random forest, gradient boosting machine and feed-forward neural network in association with the pressure at the lowest model level in the atmospheric boundary layer. Using ordinary least squares and weighted least squares regression models, Pulugurtha et al. [51] investigated the influence of contributing factors and predicting visibility at road link-level (small domain) in the state of North Carolina (USA), and they found that elevation and cloud cover are negatively associated with low visibility. Therefore, one can deduce that the ML algorithms could be geographically or climatologically dependent on the region where they are used.

The temporal distribution, as a function of the forecast hours, of the error issued from the developed models for estimating visibility over the study domain, shows that the performance of the tree-based algorithms and the neural network-based models have similar behaviour except for the generalized linear model, which confirms the worst performance for all forecast hours. Some discrepancies are noticed in the times series of mean bias, mean absolute error and mean absolute relative error for some machine-learning algorithms. It is demonstrated that it is due to the disproportionality of the number of synoptic stations operating during the whole day and that do not operate during the night. This also could be explained by the scarcity of reduced visibilities below 5 km and the prevalence

of good visibility during daytime, and the uncertainties in the mesoscale model forecast associated with the model error. One additional factor could be the uncertainties included in the observed horizontal visibility, particularly during night-time, estimated by the observer using landmarks around weather stations [47].

5 Conclusion

Three subclasses of machine-learning regression techniques (feed-forward neural network, generalized linear model and tree-based ensemble methods) have been used in this research work to diagnose horizontal visibility from a set of meteorological parameters issued from an operational mesoscale NWP model at the Moroccan Weather Service (AROME with 2.5 km of horizontal grid and 60 vertical levels). In the short-range forecasting framework, the proposed machine-learning algorithms were derived for hourly forecasts through a data-mining process up to 24 h. Their performance has been assessed against the persistence as a benchmark, over the continental part only, based on real data collected at 36 synoptic stations over 2 years, spatially distributed over topographically heterogeneous terrain. To take advantage of all the data of all the locations to estimate the horizontal visibility, the data for each location were individually split into two subsets ensuring that all months, hours, reduced and good visibilities were well represented at 70% in training and at 30% in testing data sets.

The tree-based ensemble methods have shown some improvement in visibility forecasting in comparison with the operational visibility diagnostic scheme based on Kunkel's formula and also with persistence. In fact, it is shown that XGB (extreme gradient boosting) is a suitable tool for estimating visibility over a large domain. When applied to an individual station, XGB model performance has high correlation and lowest error over the used synoptic network. In addition, it is demonstrated that ML algorithms perform better when the forecast depends on multiple predictors instead of only a few with very high importance. Regarding the robustness of the algorithms to the features correlation, it is found that their performance decreases when principal components are used instead of raw correlated data. In addition, their performance is very sensitive to disproportionality of data availability between daytime and night-time data sets.

In summary, the tree-based ensemble methods have shown some benefit in visibility diagnosing in comparison with the operational visibility diagnostic scheme. However, these techniques have shown some limitations. One possible improvement for the machine-learning techniques performance would be a rapid update cycle which takes

into account the latest observation and the previous forecast error. On the other hand, more investigation must be done to improve the ML-developed models by optimizing the architecture for FFN network and the hyperparameters tuning for the tree-based ensemble methods. This study has addressed the possibility of using machine-learning techniques as a benchmark to diagnose visibility from NWP kilometric model outputs by developing a unified model. Perhaps, it would be better to try developing many submodels dealing with regional features.

Acknowledgements Dr. Siham Sbii and Dr. Mohamed Nabhani from the Moroccan Weather Service are acknowledged for the fruitful discussions during this study and for their comments on an early version of the manuscript.

Author Contributions DB conceived the experimental framework; AO performed the data preparation and numerical computation; DB and AO contributed to the results analysis; DB wrote the manuscript with inputs from AO.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Gultepe I, Sharman R, Williams PD, Zhou B, Ellrod G, Minnis P, Trier S, Griffin S, Yum SS, Gharabaghi B et al (2019) A review of high impact weather for aviation meteorology. *Pure Appl Geophys* 176(5):1869–1921
2. Shan Y, Zhang R, Gultepe I, Zhang Y, Li M, Wang Y (2019) Gridded visibility products over marine environments based on artificial neural network analysis. *Appl Sci* 9(21):4487
3. Kutty SG, Agnihotri G, Dimri A, Gultepe I (2019) Fog occurrence and associated meteorological factors over Kempegowda International Airport, India. *Pure Appl Geophys* 176(5):2179–2190
4. Bari D, Bergot T (2018) Influence of environmental conditions on forecasting of an advection-radiation fog: a case study from the Casablanca region, Morocco. *Aerosol Air Qual Res* 18(1):62–78
5. Bari D, Bergot T, El Khlifi M (2015) Numerical study of a coastal fog event over Casablanca, Morocco. *Q J R Meteorol Soc* 141(690):1894–1905
6. Koračin D, Dorman CE, Lewis JM, Hudson JG, Wilcox EM, Torregrosa A (2014) Marine fog: a review. *Atmos Res* 143:142–175
7. Vislocky RL, Fritsch JM (1997) An automated, observations-based system for short-term prediction of ceiling and visibility. *Weather Forecast* 12(1):31–43
8. Hansen B (2007) A fuzzy logic-based analog forecasting system for ceiling and visibility. *Weather Forecast* 22(6):1319–1330
9. Kutty SG, Dimri AP, Gultepe I (2019) Climatic trends in fog occurrence over the indo-gangetic plains. *Int J Climatol*
10. Kunkel BA (1984) Parameterization of droplet terminal velocity and extinction coefficient in fog models. *J Clim Appl Meteorol* 23(1):34–41
11. Bari D (2019) A preliminary impact study of wind on assimilation and forecast systems into the one-dimensional fog forecasting model COBEL-ISBA over Morocco. *Atmosphere* 10(10):615

12. Lafore JP, Stein J, Asencio N, Bougeault P, Ducrocq V, Duron J, Fischer C, Hèreil P, Mascart P, Masson V et al (1997) The meso-NH atmospheric simulation system. Part I: adiabatic formulation and control simulations. *Ann Geophys* 16:90–109
13. Gultepe I, Müller MD, Boybeyi Z (2006) A new visibility parameterization for warm-fog applications in numerical weather prediction models. *J Appl Meteorol Climatol* 45(11):1469–1480
14. Claxton B (2008) Using a neural network to benchmark a diagnostic parametrization: the Met Office's visibility scheme. *Q J R Meteorol Soc* 134(635):1527–1537
15. Bari D, El Khilfi M (2015) LVP conditions at Mohamed V airport, Morocco: local characteristics and prediction using neural networks. *Int J Basic Appl Sci* 4(4):354–363
16. Bartoková I, Bott A, Bartok J, Gera M (2015) Fog prediction for road traffic safety in a coastal desert region: improvement of nowcasting skills by the machine-learning approach. *Bound-Layer Meteorol* 157(3):501–516
17. Marzban C, Leyton S, Colman B (2007) Ceiling and visibility forecasts via neural networks. *Weather Forecast* 22(3):466–479
18. Glahn B, Schnapp AD, Ghirardelli JE, Im JS (2017) A LAMP-HRRR Meld for improved aviation guidance. *Weather Forecast* 32(2):391–405
19. Cornejo-Bueno L, Casanova-Mateo C, Sanz-Justo J, Cerro-Prada E, Salcedo-Sanz S (2017) Efficient prediction of low-visibility events at airports using machine-learning regression. *Bound-Layer Meteorol* 165(2):349–370
20. Kneringer P, Dietz SJ, Mayr GJ, Zeileis A (2018) Probabilistic nowcasting of low-visibility procedure states at Vienna International Airport during cold season. *Pure Appl Geophys* 176:2165–2177
21. Bari D (2018) Visibility prediction based on kilometric NWP model outputs using machine-learning regression. In: 2018 IEEE 14th international conference on e-Science (e-Science), pp 278–278
22. Tardif R, Rasmussen RM (2008) Process-oriented analysis of environmental conditions associated with precipitation fog events in the New York City region. *J Appl Meteorol Climatol* 47(6):1681–1703
23. Seity Y, Brousseau P, Malardel S, Hello G, Bénard P, Bouttier F, Lac C, Masson V (2011) The AROME-France convective-scale operational model. *Mon Weather Rev* 139(3):976–991
24. Pinty JP, Jabouille P (1998) A mixed-phased cloud parameterization for use in a mesoscale non-hydrostatic model: simulations of a squall line and of orographic precipitation. In: Conference on cloud physics: 14th conference on planned and inadvertent weather modification, pp 17–21
25. Morcrette JJ (1991) Radiation and cloud radiative properties in the European Centre for Medium Range Weather Forecasts forecasting system. *J Geophys Res Atmos* 96(D5):9121–9132
26. Cuxart J, Bougeault P, Redelsperger JL (2000) A turbulence scheme allowing for mesoscale and large-eddy simulations. *Q J R Meteorol Soc* 126(562):1–30
27. Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. *Int J Forecast* 14(1):35–62
28. Nelder JA, Wedderburn RWM (1972) Generalized linear models. *J R Stat Soc Ser A (Gen)* 135(3):370–384
29. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
30. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, <http://www.deeplearningbook.org>
31. Zhou ZH (2012) Ensemble methods: foundations and algorithms. Chapman and Hall/CRC, Boca Raton
32. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer, Berlin
33. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 785–794
34. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. *Front Neurobot* 7:21
35. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
36. Parikh CR, Pont MJ, Jones NB (1999) Improving the performance of multi-layer perceptrons where limited training data are available for some classes. In: Proceedings of ninth international conference on artificial neural networks, Institution of Electrical Engineers, Edinburgh, UK, vol 1, pp 227–232
37. Fabbian D, de Dear R, Lelleyett S (2007) Application of artificial neural network forecasts to predict fog at Canberra International Airport. *Weather Forecast* 22(2):372–381
38. Gurney K (2014) An introduction to neural networks. CRC Press, Boca Raton
39. Hecht-Nielsen R (1987) Kolmogorov's mapping neural network existence theorem. In: Proceedings of the IEEE international conference on neural networks III, IEEE Press, pp 11–13
40. Zhang C, Wu M, Chen J, Chen K, Zhang C, Xie C, Huang B, He Z (2019) Weather visibility prediction based on multimodal fusion. *IEEE Access* 7:74776–74786
41. Dueben PD, Bauer P (2018) Challenges and design choices for global weather and climate models based on machine learning. *Geosci Model Dev* 11(10):3999–4009
42. Jolliffe IT (1990) Principal component analysis: a beginner's guide-I. Introduction and application. *Weather* 45(10):375–382
43. Abdi H, Williams LJ (2010) Principal component analysis. *Wiley Interdiscip Rev Comput Stat* 2(4):433–459
44. Barry RG, Carleton AM (2001) Synoptic and dynamic climatology. Psychology Press, New York
45. Jolliffe I (2011) Principal component analysis. In: Lovric M (ed) International encyclopedia of statistical science. Springer, Berlin, pp 1094–1096
46. Steyn A (1994) Modern statistics in practice. Van Schaik, Bellville
47. WMO (2014) Guide to meteorological observing and information distribution systems for aviation weather services. WMO-No. 731. World Meteorological Organization, Geneva
48. Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine-learning. ACM, pp 161–168
49. Gultepe I, Tardif R, Michaelides S, Cermak J, Bott A, Bendix J, Müller MD, Pagowski M, Hansen B, Ellrod G et al (2007) Fog research: a review of past achievements and future perspectives. *Pure Appl Geophys* 164(6–7):1121–1159
50. Hsieh WW, Tang B (1998) Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull Am Meteorol Soc* 79(9):1855–1870
51. Pulugurtha SS, Mane AS, Duddu VR, Godfrey CM (2019) Investigating the influence of contributing factors and predicting visibility at road link-level. *Heliyon* 5(7):e02105

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.