Atmospheric Research 295 (2023) 106991

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

Efficient prediction of fog-related low-visibility events with Machine Learning and evolutionary algorithms

C. Peláez-Rodríguez [a,*], J. Pérez-Aracil [a], C. Casanova-Mateo [b], S. Salcedo-Sanz [a]

[a] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares 28805, Spain
[b] Department of Computer Systems Engineering, Universidad Politécnica de Madrid, Madrid 28038, Spain

ARTICLE INFO

Keywords:
Low-visibility events
Orographic fog
Machine Learning algorithms
Reanalysis data
Evolutionary optimization algorithms
Iterative forward selection

ABSTRACT

Low visibility events are a severe problem for road transport, causing accidents and major economic losses. Their accurate prediction may help prevent these problems. For that purpose, machine and deep learning techniques have been applied for fog prediction using in situ meteorological data and persistence variables as baseline predictors. These techniques have been evaluated for different prediction time-horizons: 1 h, 3 h and 6 h. The effect of including data extracted from ERA5 Reanalysis as predictive variables has been studied. A database, covering 23 months, has been used, which contains visibility and other meteorological variables measured in Mondoñedo, Galicia, Spain. A 222000 km^2 region around Mondoñedo has been delimited. Thus, a proposed iterative forward selection algorithm based on evolutionary algorithms has been applied to determine the optimal variables and nodes in the region for each regressor model. Both Differential Evolution and Particle Swarm Optimization have been used as optimization algorithms, and an improvement of up to 17.3% with respect to the baseline databases have been obtained. Finally, an analysis of the most frequently selected variables by the evolutionary algorithms has been conducted, leading us to conclude which variables and geographical nodes provide better information to the prediction models.

1. Introduction

Fog remains as one of the meteorological phenomena with the greatest impact on weather-dependent human activities (Torres-López et al., 2022), mainly agriculture (Baldocchi and Waller, 2014) and transportation (Durán-Rosal et al., 2018; Guerreiro et al., 2020; Fabbian et al., 2007), among others. Focusing on road transport, low-visibility conditions due to dense fog, mist, or their combination, provide highly unsafe scenarios on roads, presenting numerous challenges to drivers and increasing the risk for passengers, including strong visual effects, decreasing the ability to see through the air (Palvanov and Cho, 2019). They also reduce the transport capacity on affected roads, leading to closures in extreme cases, with the consequent economic cost (Cho and Kim, 2005).

This paper deals with specific extreme low-visibility events caused by orographic fog in Spain. Specifically located in the North of Spain, the A-8 motor-road is the largest and most important highway running along the Cantabrian coast. It runs from the Basque Country to the province of Lugo in Galicia (North-West Spain). There, the specific geographic location, together with the surrounding atmosphere conditions, favors the appearance of a specific type of fog called *orographic fog* (Wilson and Barros, 2017). Orographic fog occurs when an air mass is forced from a low elevation to a higher elevation as it moves over rising terrain (Stull, 2015). As the air mass gains altitude, it rapidly cools adiabatically, which may raise the relative humidity up to 100%, creating clouds and causing low-visibility events if these clouds are close or in contact with the ground. This problem continues to be an unresolved issue since the construction of the highway some years ago, resulting in over 700 h per year of closures due to low-visibility events, along with a high number of traffic crashes associated with these events. The statistical analysis of low-visibility events in Mondoñedo was previously studied in Cornejo-Bueno et al. (2021). As part of the solution to this problem, the most possible accurate and reliable meteorological information is needed, in order to anticipate extreme low visibility events and therefore minimize both economic losses and the risk of traffic accidents.

Different approaches and techniques can be found in the literature for fog prediction and their associated low-visibility events. For example numerical weather prediction (NWP) models as one most widely used methods (Bergot and Guedalia, 1994; Gultepe et al., 2007; Fernández-

* Corresponding author.
E-mail address: cesar.pelaez@uah.es (C. Peláez-Rodríguez).

https://doi.org/10.1016/j.atmosres.2023.106991
Received 15 September 2022; Received in revised form 23 August 2023; Accepted 29 August 2023
Available online 8 September 2023

González et al., 2019). Several recent works evaluate the performance of mesoscale (or high-resolution) NWP models to accurately forecast fog events (Román-Cascón et al., 2016; Steeneveld et al., 2015). However, due to the local nature of this type of phenomena, forecasting low visibility events by NWP is a complex process. This is partly because fog formation is very sensitive to small-scale variations in atmospheric variables (changes in wind or in atmospheric stability) and many current models do not capture such spatial resolution (Cornejo-Bueno et al., 2021). Alternatively, data-based statistical approaches based on the processing of measured variables have been widely employed. Advances in last decades within this area provide opportunities to explore the use of Machine Learning (ML) and Deep Learning (DL) approaches to address the visibility distance forecasting problem, which is a nonlinear complex problem characterized by temporal and spatial variability.

In this regard, diverse approaches have been implemented: For example, Bayesian decision algorithms were applied in Boneh et al. (2015) to forecast fog in Melbourne Airport. A variety of shallow ML methods have been successfully implemented to solve the problem of low visibility predictions, both for classification and regression problems: in Castillo-Botón et al. (2022) shallow ML classification and regression algorithms are used to forecast the orographic fog in the A-8 motor road, Spain. In Bari and Ouagabi (2020) different ML models (tree-based ensemble, feed-forward neural network and generalized linear methods) are applied in a problem of horizontal visibility prediction, using hourly observed data at 36 synoptic land stations over the northern part of Morocco. In Kim et al. (2022) different ML approaches such as Random Forest or Neural Networks are applied to atmospheric visibility prediction in Korea. In Dietz et al. (2019) tree-based statistical models based on highly-resolved meteorological observations are applied to forecast low visibility at airports. In Zhang et al. (2022), Ortega et al. (2019) ML techniques are applied to classify low-visibility events in Chengdu (China) and Florida (USA), respectively. In Wen et al. (2023) a number of ML approaches such as Support Vector Regression (SVR), k-nearest neighbours or Random Forest have been applied to a problem of haze pollution with visibility reduction prediction in China. Fuzzy logic and neural networks has also been implemented for predictions of cloud ceiling height and horizontal visibility at airports in Hansen (2007) and Marzban et al. (2007), respectively. Somehow related to that approach, in Peláez-Rodríguez et al. (2023) an approach based on the development of inductive and evolutionary-based decision rules is proposed to exploit the explainability of the decision rules in a problem of low-visibility events prediction. Other ML approaches to estimate atmospheric visibility, including algorithms such as neural network regressors, tree-based ensembles or Random Forest approaches, among others, have been presented in recent papers (Choi et al., 2022; Cornejo-Bueno et al., 2021; Bari and Ouagabi, 2020; Yu et al., 2021; Zhai et al., 2023).

The application of different DL methods that enable the processing of a large volume of time series data has also been recently proposed. In Ortega et al. (2022) various DL methods were applied using climatological data for visibility forecasting. In Zhu et al. (2017) Deep Neural Networks (DNN) were implemented to forecast airport visibility. Convolutional Neural Networks and Auto Regressive Recurrent Neural Network (ARRNN) have also been applied in Palvanov and Cho (2019) and Jonnalagadda and Hashemi (2020), respectively. In Zang et al. (2023) a deep Recurrent Neural Network (RNN) was proposed for a problem of atmospheric visibility prediction from ground-based monitoring data in Southern China. The results obtained showed that the RNN was able to improve the results of alternative DL approaches such as CNN networks. In Peláez-Rodríguez et al. (2023) the performance of different DL ensembles was discussed in a problem of fog-related low-visibility events prediction. The ensemble approaches showed excellent performance, improving that of single DL algorithms. Finally, alternative ML models including evolutionary-based architectures were also implemented to forecast fog and low-visibility event in airports in Durán-Rosal et al. (2018).

The approach proposed in this paper is based on the use of meteorological variables from Reanalysis (together with in situ observations) to improve the prediction of ML algorithms in low-visibility events due to orographic fog. Reanalysis data are highly reliable and have already been used to accurately forecast fog (Herman and Schumacher, 2016; Zhang et al., 2022) and other meteorological variables as solar radiation (Ghimire et al., 2018) or wind speed (Salcedo-Sanz et al., 2014) before. Nevertheless, the vast domain of possibilities available when selecting the most significant meteorological variables and its geographic locations, make necessary the use of feature selection algorithms (Salcedo-Sanz et al., 2018). Moreover, the application of evolutionary methods to select geographical nodes have been extensively used in the literature in recent years to solve meteorological prediction problems (Seo et al., 2014; Salcedo-Sanz et al., 2014; Jeong et al., 2015; Salcedo-Sanz et al., 2018). Thus, In this paper, a novel iterative evolutionary forward selection algorithm is proposed to select the reanalysis variables which are most significant for the problem addressed. Once the initial database is built from data measured at Mondoñedo station and including data related to fog persistence (since this represents a significant characteristic of these type of low visibility orographic events (Salcedo-Sanz et al., 2021)), evolutionary algorithms are iteratively run to select the best reanalysis variable that minimizes the error in the validation data. At each iteration, the chosen variable is added to the database and the evolutionary algorithm is re-run. This process is repeated until no improvement in the validation error is obtained. In order to assess the performance of the proposed method, three different evolutionary algorithms (Random Search (RS), Differential Evolution (DE) and Particle Swarm Optimization (PSO)) were used to perform the search for optimal variables. In addition, regression is carried out with five different techniques, from linear regression, shallow ML methods, and neural networks, both high training-speed randomized neural networks and deep architectures. Besides, different time horizons have been used to perform the visibility forecasting (+1 h, +3 h and +6 h), so we can analyze the performance of the proposed prediction approaches with feature selection at different prediction time-horizons.

The rest of the manuscript has been organised as follows: Section 2 details the proposed methodology along with the materials used. Section 3 presents the results obtained. Finally, some discussions and conclusions are provided in Sections 4 and 5, respectively.

2. Materials and methods

This section details the proposed methodology with a description of the complete architecture (Section 2.1) along with the regression methods (Section 2.2) and the optimization algorithms (Section 2.3) used.

2.1. Iterative evolutionary forward selection algorithm

The architecture of the implemented methodology is shown in Fig. 1. The preliminary procedure comprises the formation of the initial database, which in this case includes the measured variables and those associated with visibility at previous hours, and its division into training, validation and test subsets. Training data are used to perform the training of the regression models, validation data are employed by the evolutionary algorithm for the computation of the fitness function, and test data are reserved for the evaluation of the solutions provided by the algorithm.

Then, iteratively the evolutionary algorithm is executed to search for the reanalysis variable within the selected spatial domain that provides the most improved prediction on the validation data. A new dataset is formed subsequently with the addition of the optimum reanalysis variable and the search is started again for the next variable. This process is repeated until no improvement is found in ten consecutive iterations.

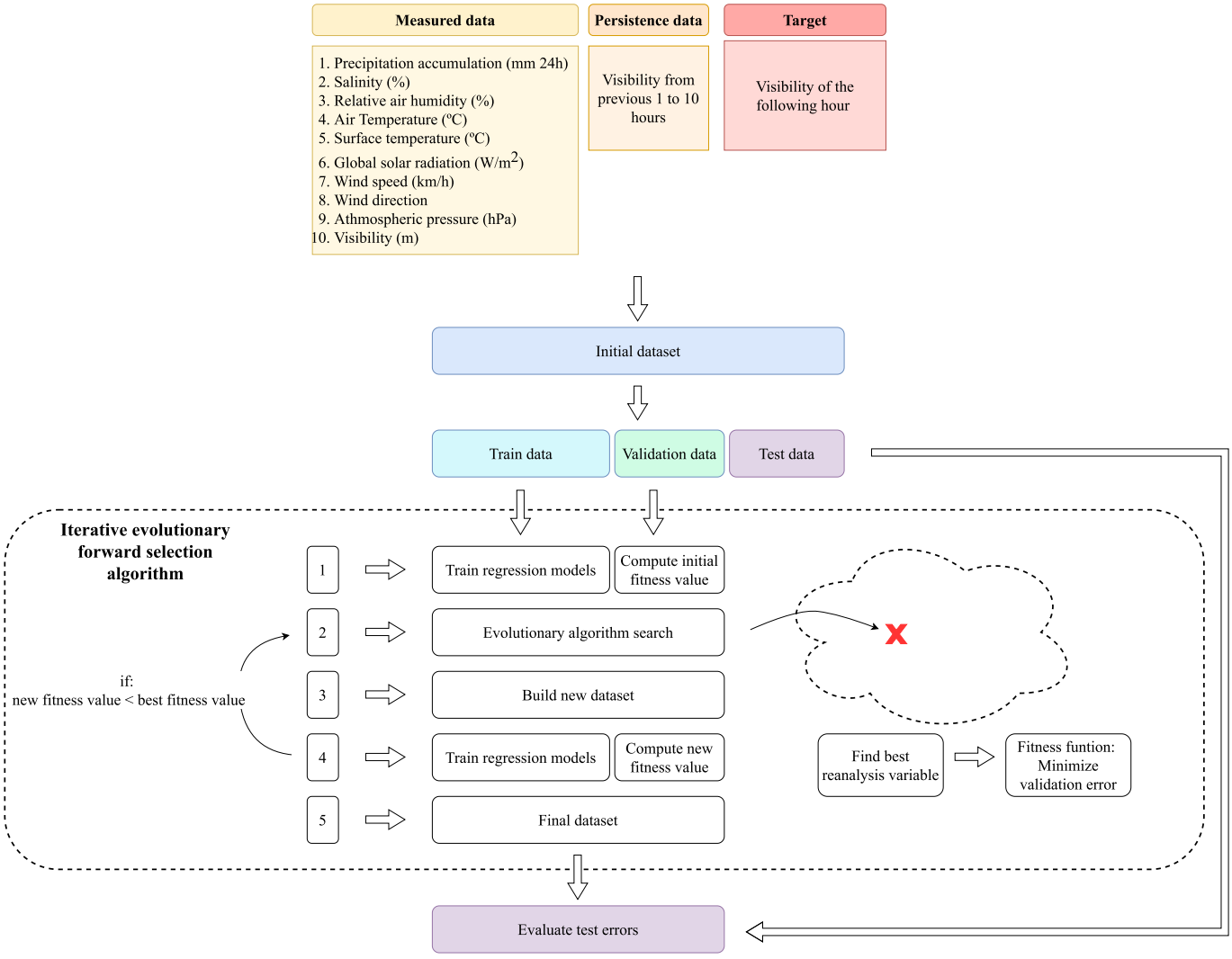

Fig. 1. Iterative evolutionary forward selection architecture.

2.2. Regression methods

Five different regression methods have been selected for performing the visibility forecasting. They include classical shallow ML methods such as Linear Regression (Section 2.2.1), Regression Trees (Section 2.2.2) and Random Forest (Section 2.2.3), and two different architectures of artificial neural networks: Extreme Learning Machine (Section 2.2.4) and Fully Connected Deep Neural Network (Section 2.2.5). At this point, it is important to point out that, since we are handling the use of evolutionary algorithms, the training speed of the regressors is a crucial aspect in its choice, as the training of these models is performed hundreds of times in each iteration. Therefore, fast-training regression models have been selected in this work (Saavedra-Moreno et al., 2013).

2.2.1. Linear regression

Linear regression (LR) (Draper and Smith, 1998) is a statistical method for modeling the linear relationship between independent variables (predictive variables) and dependent variables (target). The general formula for multiple regression models is the following:

$$Y = \beta_0 + \sum_{j=1}^{n} \beta_j X_j + \epsilon \tag{1}$$

where Y denotes the target variable; β_0 represents a bias constant; X_j is the value of the predictive variable j; β_j denotes the regression coefficient ($j = 1, 2, \ldots c, n$); and ϵ is an error term.

In the implementation of this method, the straightforward and deterministic ordinary least squares (OLS) (Dismuke and Lindrooth, 2006) method has been selected for obtaining the LR coefficients.

2.2.2. Regression Trees

Classification and Regression Trees (CART) (Loh, 2011) are ML methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Regression Trees (RT) take continuous or ordered discrete values for dependent variables, and a regression model is fitted to each node to give the predicted values of the target variable.

2.2.3. Random Forest

Random Forest (RF) (Breiman, 2001) is among the most renowned bagging-like techniques for classification and regression problems. Bagging are the simpler ensemble technique to train multiple learners and provide an unified output. It considers an ensemble composed by learners with equal architecture, that is, with same topology, number of input–output variables and parameters (Fig. 2).

RF employs decision or regression trees as predictors in a way that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Therefore

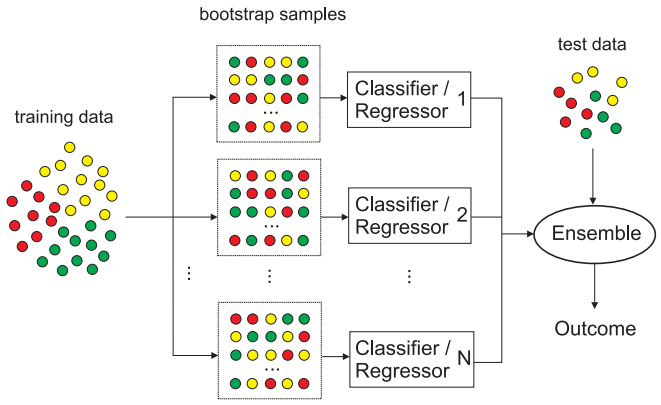

Fig. 2. Bagging technique for classification or regression problems.

RF differs from the pure bagging technique in the topology of the trees changes among them. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.

2.2.4. ELM

ELM regression method (Huang et al., 2006), is a fast-training randomized neural network, it consists of a specific architecture of Multilayer Perceptron (MLP) (Gardner and Dorling, 1998) neural network. The ELM is based on the concept that if the MLP input weights are fixed to random values, the MLP can be considered as a linear system and the output weights can be easily obtained using the pseudo-inverse of the hidden neurons outputs matrix H for a given training set (Huang et al., 2011). Algorithm 1 shows a visual summary of the ELM implementation.

Algorithm 1 Extreme Learning Machine (ELM)

Given a training set $D = \{(\mathbf{x}_i, \mathbf{t}_i) | \mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m, \mathbf{i} = 1, \ldots, N\}$,
an activation function f and an hidden neuron number M,
1:Assign arbitrary input weights $\mathbf{w}_j$ and biases $\mathbf{b}_j, j = 1, \ldots, M$
2:Compute the hidden layer output matrix $\mathbf{H}$,
where $\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_M, b_1, \ldots, b_M, \mathbf{x}_1, \ldots, \mathbf{x}_N) =$

$$\begin{bmatrix} f(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & f(\mathbf{w}_M \cdot \mathbf{x}_1 + b_M) \\ \vdots & \cdots & \vdots \\ f(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & f(\mathbf{w}_M \cdot \mathbf{x}_N + b_M) \end{bmatrix}_{N \times M}$$

3:Calculate the output weight matrix $\mathbf{B} = \mathbf{H}^{\ddagger}\mathbf{T}$,
where $\mathbf{B} = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_M^T \end{bmatrix}_{M \times m}$, $\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix}_{N \times m}$,
and $\mathbf{H}^{\ddagger}$ represents the pseudo-inverse of $\mathbf{H}$

2.2.5. Fully Connected Deep Neural Networks

Fully Connected Deep Neural Networks (FCDNN) (Alom et al., 2019) refers to a neural network architecture composed of multiple layers: an input layer, one or more hidden layers, and one output layer. Each layer is composed of units called nodes or neurons which are connected to every other node in the neighboring layers by means of weighted links. The representation of this kind of network as a directed acyclic graph is shown in Fig. 3.

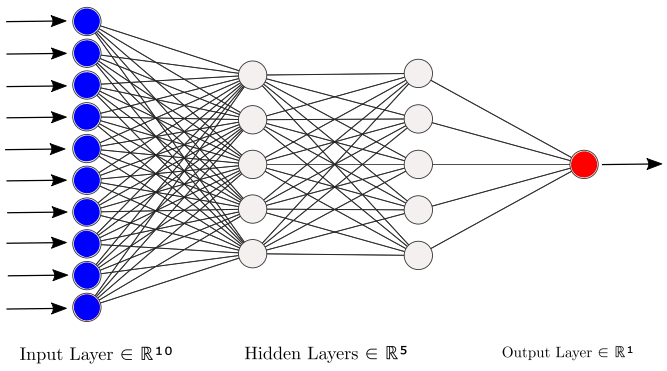

Fig. 3. Structure of a FCDNN.

2.3. Optimization algorithms

Three different optimization algorithms have been employed to accomplish the validation of the feature selection framework proposed. First, a RS algorithm has been established as a baseline (Section 2.3.1). Then, two well known state-of-the-art evolutionary algorithms have been implemented: DE (Section 2.3.2) and PSO (Section 2.3.3).

2.3.1. Random search

A random search algorithm has been tested as a baseline for assessing the potential of the proposed methodology. For this, at each iteration of the forward selection algorithm, a meteorological variable from Reanalysis is randomly selected, and subsequently the validation error is evaluated at 50 random geographic nodes within the search space. The best performing node in the fitness function is selected as the optimum for that iteration, and in the case of an improvement over the best absolute error obtained so far, that reanalysis variable at the optimum node is added to the dataset. As shown in Fig. 1, this procedure is repeated until there are no improvements during 10 consecutive iterations.

2.3.2. Differential Evolution

DE (Storn and Price, 1997) consists of a stochastic, population-based and direct searching method designed for global optimization problems. DE uses real encoding method, its survival strategy based on the simple differential mutation operation and one-to-one competition reduces the complexity of the genetic operation. Therefore, it is suitable to solve the optimization problem in the complex environment, which can not be solved using conventional mathematical programming methods. It keeps a population with N_p individuals, where every individual within the population stands for a possible solution to the problem. Individuals are represented by a vector $X_{i,g}$, where $i = 1, \ldots, N_p$ and g refers to the index of the generation. A DE cycle consists of three consecutive steps: mutation, crossover and selection, which are described as follows.

Mutation is carried out to generate random perturbations on the population. For each individual, a mutant vector is generated. Different approaches for DE mutation are present in literature (Price et al., 2006). "Best" mutation strategy has been selected in this paper: it attempts to mutate the best individual of the population, according to Eq. (2), where $V_{i,g}$ denotes the mutated vector, i is the index of the vector, g stands for the generation index, $r_1, r_2 \in 1, \ldots, N_p$ are randomly created integers, $X_{best,g}$ denotes the best solution in the population and F is the scaling factor in the interval $[0, 2]$. This mutation strategy uses the scaled difference between two randomly selected vectors to mutate the best individual in the population. Fig. 4 shows how a new mutant vector is obtained with this strategy, where d denotes the difference vector between $X_{r1,g}$ and $X_{r2,g}$.

$$V_{i,g} = X_{best,g} + F \cdot (X_{r1,g} - X_{r2,g}) \tag{2}$$

Crossover carries out the combination of every individual with the mutant vector created in mutation stage. The new solutions created are called trial vectors and are denoted by $T_{i,g}$ for individual i at generation g. Every parameter in the trial vector are decided following Eq. (3), where j represent the index of every parameter in a vector, CR is the probability of recombination, and J_{rand} denotes a randomly selected integer within $(1, \ldots, N_p)$ to ensure that at least one parameter from mutant vector enters the trial vector

$$T_{i,g}[j] = \begin{cases} V_{i,g}[j] & \text{if} \quad rand[0,1] < CR \quad \text{or} \quad j = j_{rand} \\ X_{i,g}[j] & \text{otherwise} \end{cases} \tag{3}$$

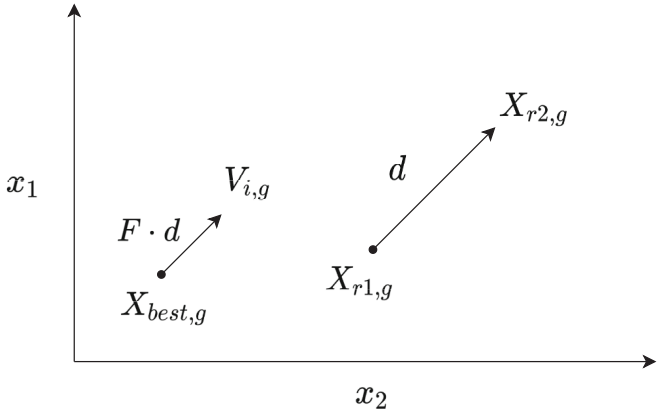

Fig. 4. Best mutation with one difference vector.

Lastly, the selection consists of comparing each trial vector with its parent solution and further decides the winner to survive into the next generation. The individuals for the new generation are chosen following Eq. (4), where $T_{i,g}$ is the trial vector, $X_{i,g}$ the individual in the population, $X_{i,g+1}$ denotes the individual in the next generation and $f()$ represents the fitness value of the corresponding individual.

$$X_{i,g+1} = \begin{cases} T_{i,g} & \text{if } f(T_{i,g}) < f(X_{i,g}) \\ X_{i,g} & \text{otherwise} \end{cases} \tag{4}$$

DE algorithm was implemented following the pseudo-code showed in the Algorithm 2 and the differential evolution function from Python library scipy.optimize was used.

Algorithm 2 Differential evolution algorithm

Initialize variables:
Crossover Rate (CR)
Mutation Rate (F)
Population Size (PS)
Maximum Iterations (maxIter)
Bounds ⇒ (LB,UB)
Initialize the population with randomly created individuals
Compute fitness value for all individuals
for iter **to** maxIter **do**
  Create mutant vector (Eq. (2))
  Create trial vectors (Eq. (3))
  Compute fitness values of trials vectors.
  Select winning vectors for the next generation (Eq. (4))
end for

2.3.3. Particle Swarm Optimization

PSO (Kennedy and Eberhart, 1995) is an evolutionary method that exploits simple analogues of social interaction rather than purely individual cognitive abilities (Poli et al., 2007). A number of particles is placed in the search space, each meaning a solution for the problem; therefore, objective function is evaluated at its current location. Then, each particle determines its movement through the search space by combining some aspect of the history of its own current and best locations with those of some neighbor members of the swarm, with some random perturbations. After all particles have been moved, the next iteration begins. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function.

Each individual in the particle swarm is composed of three D-dimensional vectors, where D is the dimensionality of the search space. These vectors are x_i, p_i and v_i, which represent the current position of the particle i, its historical best position and its velocity, respectively.

The current position x_i may be considered as a set of coordinates describing a point in the search space. On each iteration of the algorithm, the current position is evaluated as a problem solution. If that position is better than any that has been found so far by the particle i, then the coordinates are stored in the p_i vector and the fitness value is saved at a variable named $pbest_i$. The velocity vector determines how the particles move between iterations by adding v_i coordinates to x_i (Eq. (5)). The velocity of each particle is iteratively adjusted so that the particle stochastically moves according to its best historical solution and the best point found by any member of its topological neighborhood (named as particle g) (Eq. (6)), where $U(0, \phi_i)$ represents a vector of random numbers uniformly distributed in $[0, \phi_i]$ which is randomly generated at each iteration and for each particle and $\times$ denotes the component-wise multiplication.

Therefore, the particle swarm is more than just a collection of particles. A particle by itself has almost no power to solve any problem; progress occurs only when the particles interact.

$$x_i \rightarrow x_i + v_i \tag{5}$$

$$v_i \rightarrow v_i + U(0, \phi_1) \times (p_i - x_i) + U(0, \phi_2) \times (p_g - x_i) \tag{6}$$

PSO algorithm has been implemented following the pseudo-code showed in the Algorithm 3, with the Python pyswarm package.

Algorithm 3 Particle Swarm Optimization

Initialize a population array of particles with random positions and velocities.
for iter **to** maxIter **do**
  Evaluate fitness function for each particle.
  Compare particle's fitness evaluation $f(x_i)$ with its $pbest_i$
  Update $pbest_i = f(x_i)$ and $p_i = x_i$ if $f(x_i)$ is better than $pbest_i$
  Identify the particle in the neighborhood with the best $f(x_i)$, and update variable g.
  Change the velocity and position of the particle according to the Eqs. (5) and (6)
end for

3. Experiments and results

The experimental work of this paper is detailed in this section. First, a description of the time series visibility along with the specification of the predictive variables are presented in Sections3.1 and 3.2, respectively. After that, baseline results are shown in Section 3.3. Then, a sensitivity analysis is conducted in Section 3.4 to qualitatively assess the significance of using Reanalysis variables to improve visibility prediction. Subsequently, the steps carried out in the implementation of the presented algorithm are specified in Section 3.5, as well as the codification and decoding process of the optimization algorithms solutions. Finally, results for the proposed method are shown and analyzed in Section 3.6.

3.1. Data description

Real visibility data measured at the Mondoñedo weather station, Galicia, Spain (43.38°N, 7.37°W) from 1st January 2018 to 30th November 2019 have been considered as the time series target variable in this paper. The visibility data used were acquired by a Biral WS-100 visibility sensor, which operates between 0 to 2000 m. A sampling frequency of 1 h matching the reanalysis meteorological variables frequency has been selected for this study. Fig. 5 illustrates the first 500 samples of the visibility time series, where it can be easily perceived how the data tends to approach values close to 2000 or to 0, while avoiding intermediate ranges. This is further illustrated by plotting the density function of the data (Fig. 6), where two peaks may be found: one near 0 m, corresponding to very low visibility levels, and a much more prominent peak near 2000 m, corresponding to events where visibility is very good.

3.2. Predictive variables

The predictive variables used in this study may be divided into three different groups. In first place, exogenous meteorological variables registered by the same weather station at Mondoñedo have been considered to established the baseline prediction. Table 1 shows the 11 meteorological variables considered. The time horizon of the visibility forecast have been established in +1, +3 and +6 h; therefore, measured variables acquired at $t = h$ are used as predictors to forecast visibility at

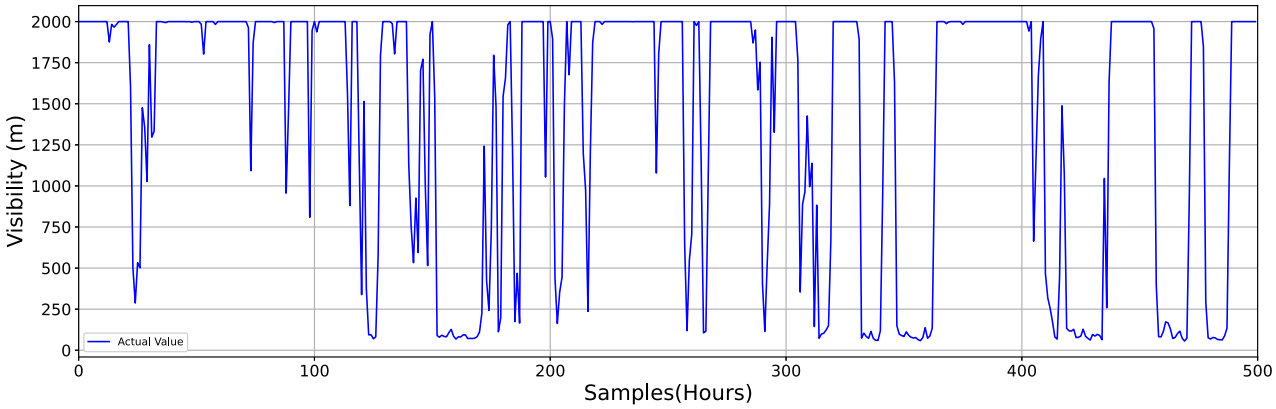

Fig. 5. First 500 samples of the time series visibility considered as target variable.

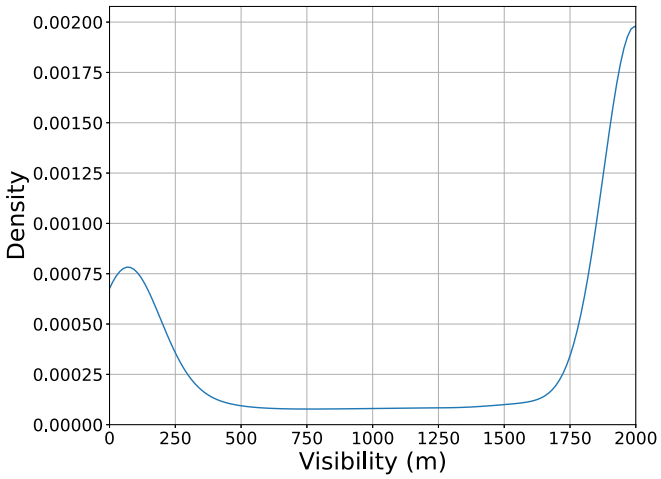

Fig. 6. Density function of the visibility time series variable.

Table 1
Predictive variables measured at Mondoñedo, Galicia, Spain.

Variable name	Measured variable
qp_Ra	Accumulated precipitation
at	Air temperature
ap	Atmospheric pressure
td	Dew temperature
st	Floor temperature
gr	Global solar radiation
hr	Relative Humidity
sa	Salinity
wd	Wind direction
ws	Wind speed
vs	Visibility

$t = h+1, t = h+3$ and $t = h+6$, respectively.

Then, considering that persistence plays an important role at the detection of these type of orographic fog, as studied in recent works as Salcedo-Sanz et al. (2021), ten additional variables representing the visibility measured at previous hours (from $t = h-1$ to $t = h-10$) have been considered, named as vs_1 to vs_10, respectively.

Fig. 7 depicts the correlation coefficients between the predictive variables (measured and persistence variables) and the target variables for the three time forecasting horizons considered, showing how it is precisely these previous hours visibility variables the ones with the highest correlation coefficients. It is also noteworthy the strong negative correlation that exists between the visibility and the dew temperature and the relative humidity, which are obviously two denoting factors for the presence of fog.

The last group of predictor variables, which are in fact the object of this paper, consists of meteorological variables extracted from ERA5 Reanalysis (Hersbach et al., 1979). This database provides very accurate historical hourly information on meteorological variables related to temperature, pressure, precipitation and snowfall among others; with a resolution of 0.25 degrees of longitude and latitude between grids. Therefore, using Reanalysis data can be considered to be equivalent to have an initial prediction from an operational model, but avoiding the intrinsic error of the prediction model.

Aiming at improving the visibility prediction obtained with the persistence variables and the meteorological variables measured at the Mondoñedo station, Reanalysis variables of temperature, pressure and different wind components at different heights have been selected to find out which geographical locations surrounding Mondoñedo and in the northwestern area of the Iberian Peninsula, (including part of the Cantabrian Sea), provide the regression models with the greatest improvement. Table 2 lists the 10 predictive variables selected per node. Fig. 8 displays the geographical area in which the optimal Reanalysis variables were searched, which encompasses approximately 660 × 330 km, representing a total of 288 nodes (24 × 12). This results in a search space of 2880 variables for the optimization algorithm.

3.3. Baseline results

First, initial visibility forecasting is carried out by using the set of meteorological variables measured at the Mondoñedo weather station as predictors (Table 1). Mean Absolute Error (MAE) has been selected as the performance metric for this regression problem (Eq. (7)).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

where $\hat{y}$ represents predicted values (provided by the proposed model) and y are the actual observed values. The subscript i is used to refer to a single sample $y_i = y[i]$.

Results for this initial prediction are shown in Table 3 for the three time horizons considered. It can be observed how the FCDNN outperforms the rest of the methods by a large margin. Also, it is noticeable the worsening of the predictions as the time horizon increases.

Next, variables related to the visibility information in the preceding hours are included in the training of the models. Table 4 exhibits the performance of the predictions made, where in all cases an improvement in the results over the initial prediction is observed. This improvement is quantified and shown in the tables (values in brackets), with an average improvement value of 2.24%.

Thus, the significance of considering these persistence variables for the resolution of this regression problem is evidenced. From this point

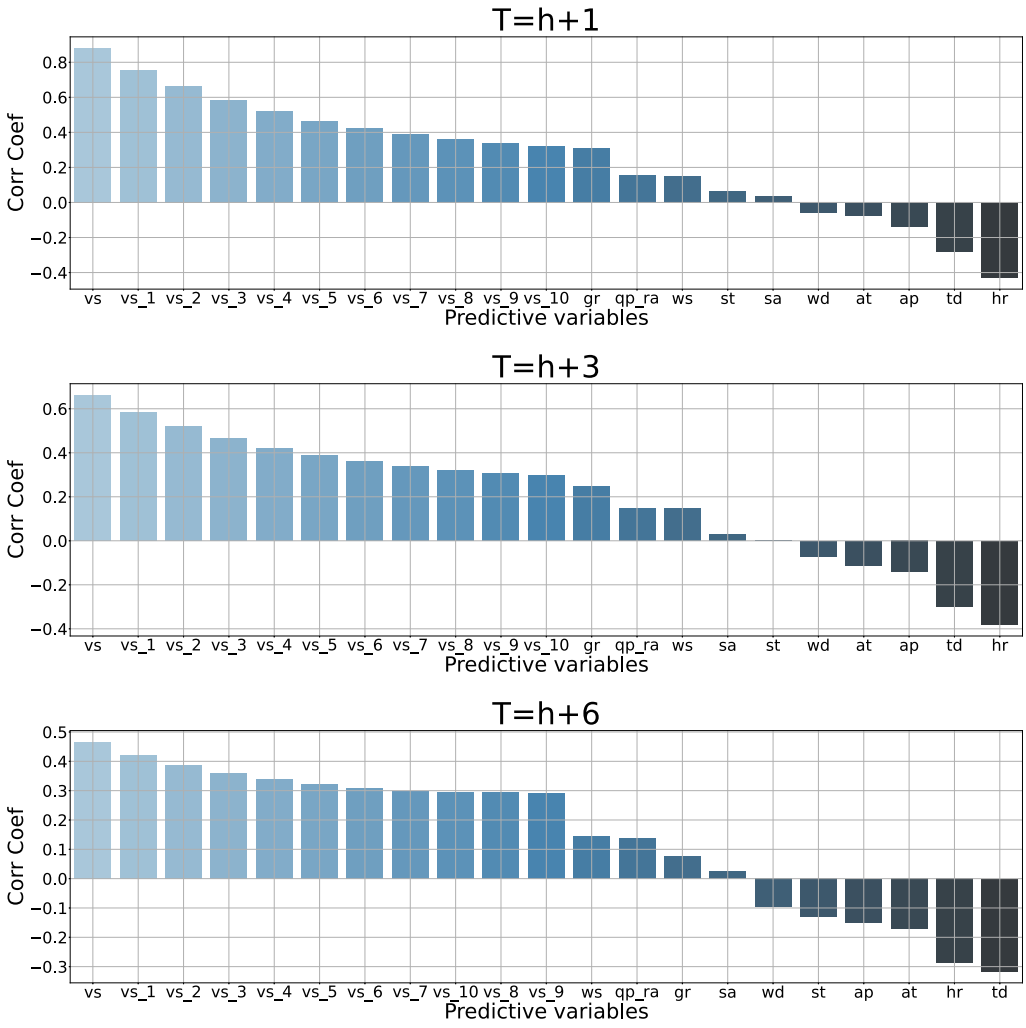

Fig. 7. Correlation coefficients between the predictive and the target variables for the three time forecasting horizons considered.

Table 2
Predictive variables considered at each node from the ERA-5 reanalysis.

Variable name	ERA5 Variable
d2m	2 m dewpoint temperature
t2m	2 m temperature
sp	Surface pressure
msl	Mean sea level pressure
u10	10 m u-component of wind
u10n	10 m u-component of neutral wind
u100	100 m u-component of wind
v10	10 m v-component of wind
v10n	10 m v-component of neutral wind
v100	100 m v-component of wind

on, these will be considered as the baseline models, whose performance will be further improved by introducing additional predictor variables extracted from Reanalysis after applying a feature selection algorithm based on evolutionary optimization methods.

3.4. Sensitivity analysis

A sensitivity analysis is conducted in this section in order to assess the effect of the location for each reanalysis variable on the prediction performance. For this purpose, for each meteorological variable of those listed in Table 2, 100 random geographic nodes have been selected within the area of study (Fig. 8). Each variable was separately added to the dataset composed by the measured variables in Mondoñedo weather station and by those related to persistence, resulting in 100 different datasets for each meteorological variable. Then, the regressors were feeded with each of the datasets and the MAE on the test data was computed.

Fig. 9 displays the sensitivity analysis for each Reanalysis variable when LR is defined as the regression method. Here, the relationship between better prediction performance (lower MAE) and the geographic location of the reanalysis nodes is observed. In these figures, a darker color denotes a lower MAE and, consequently, a better prediction performance. Additionally, the blue cross indicates the location of the Mondoñedo meteorological station where the visibility is measured.

It is possible to analyze how for the temperature related variables (d2m and t2m), the most characteristic nodes are located in the northern region of the selected area. The same can be observed for the pressure variables (sp and msl), where the best performance is obtained in the northwest region. For the wind-related variables, however, the most significant nodes are found in the west area, around Mondoñedo. In addition, comparing the best MAE found for each variable, it can be noticed that there are variables that perform better than others.

Next, Fig. 10 illustrates the same sensitivity analysis described previously using ELM as regression method. Here, a major dispersion are observed in the results, although similar conclusions can be drawn as in the previous case: the most significant pressure variables are found in the northern region, and in the western region for the wind-related variables. In the case of temperature, for instance, no pattern is

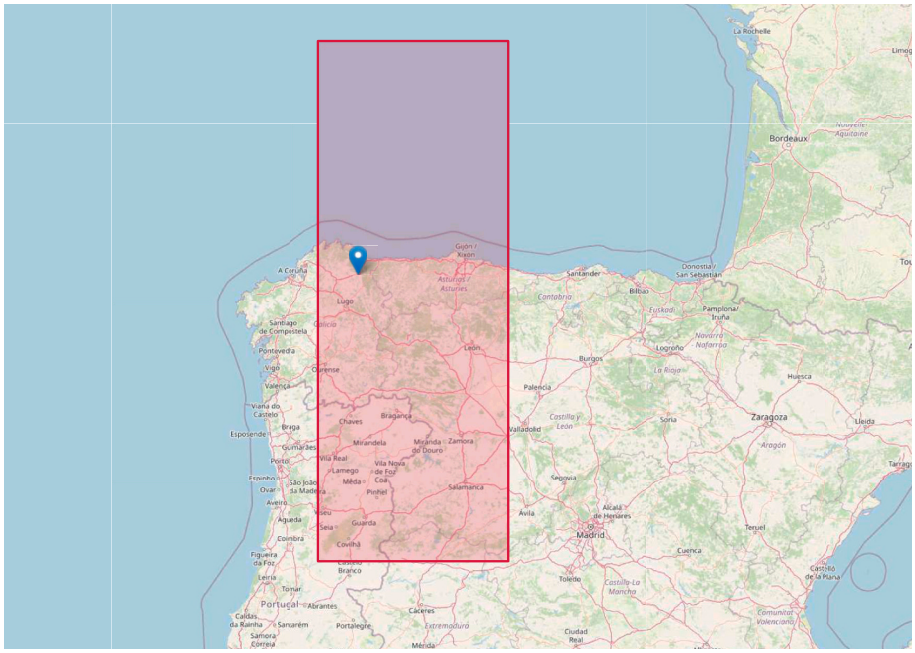

Fig. 8. Northwestern area of the Iberian Peninsula, where the Mondoñedo Station is indicated with a blue marker. The red area represents the geographical area in which optimal Reanalysis variables were searched.

Table 3
Visibility forecasting performance for three prediction time-horizons when using measured meteorological variables at Mondoñedo.

ML Method	+1 h	+3 h	+6 h
LR	240.05	476.33	628.83
RT	238.42	458.40	627.08
RF	199.35	417.21	569.25
ELM	227.25	425.71	570.33
FCDNN	**171.07**	**326.60**	**452.33**

Table 4
Visibility forecasting performance for three prediction time-horizons considering measured meteorological variables and persistence variables. In brackets the percentage improvement over the initial prediction is indicated.

ML Method	+1 h	+3 h	+6 h
LR	238.54 (0.63%)	469.81 (1.37%)	616.15 (2.02%)
RT	229.95 (3.55%)	443.68 (3.21%)	606.68 (3.25%)
RF	193.07 (3.15%)	409.75 (1.79%)	566.45 (0.49%)
ELM	221.42 (2.57%)	421.04 (1.10%)	567.29 (0.53%)
FCDNN	**161.41** (5.65%)	**322.63** (1.22%)	**438.59** (3.04%)

observed in the distribution of results.

The appearance of a higher dispersion in the distribution of the most optimal geographic nodes for each variable is something that, as it has been shown, depends on the regression method used. It can be noticed that more explainable and simpler models, which will have worse performance, show more regular patterns, while more complex models present, in general, more distributed and irregular patterns. Nonetheless, in all cases it is observed that a dependence exists between the geographical position of the nodes and the performance of the models exists, meaning that there will be some nodes that will produce a larger improvement in the results than others.

3.5. Implementation

Details on the specific implementation of the proposed algorithm are provided in this section.

In first place, the codification procedure for the solutions provided by the evolutionary algorithms is detailed in Table 5. The evolutionary algorithms output correspond to an integer value ranging from 1 to $N \times m + m$, where N is the number of geographical nodes and m denotes the number of meteorological variables considered per node. Therefore, in order to determine which meteorological variable corresponds to a given solution, we only need to calculate the remainder of the division between the solution and m. Similarly, the quotient of this division refers to the geographic node of the solution, once all the nodes have been sorted from west to east and from north to south, i.e.: node 1 corresponds to the northwest of the studied region and node N to the southeast.

In the evolutionary algorithms implementation, the following parameters have been defined: In the case of DE, PS = 10, CR = 0.7, F = 1; and in the case of PSO, Swarm size = 10. In both scenarios, the maximum number of iterations has been defined as 50 (a small number of iterations have been set since the evolutionary methods are run multiple times by the feature selection algorithms) and the search limits have been established between 1 and $N \times m + m$.

3.6. Results

This section reports the results achieved for the different evolutionary algorithms used and the three time-horizons considered (+1 h, +3 h and +6 h). The iterative feature selection algorithm was run 5 independent times for each case. The results shown correspond to the average of these 5 executions.

Tables 6–8 report the results obtained for the three prediction time-horizon considered. In them, the MAE obtained for each case is detailed, along with the improvement percentage of the model after including the reanalysis variables compared to the baseline models (Table 4). Finally, the number of reanalysis variables chosen by the feature selection algorithm is also displayed.

Several conclusions may be extracted from this tables. First, it is important to clarify that the optimization algorithms are launched with the aim of minimizing the error obtained in the prediction of the validation data, while the tables show the errors in the test data. Thus, it is possible that the addition of some reanalysis variable may lead to a small improvement in the validation data but not in the test data, which

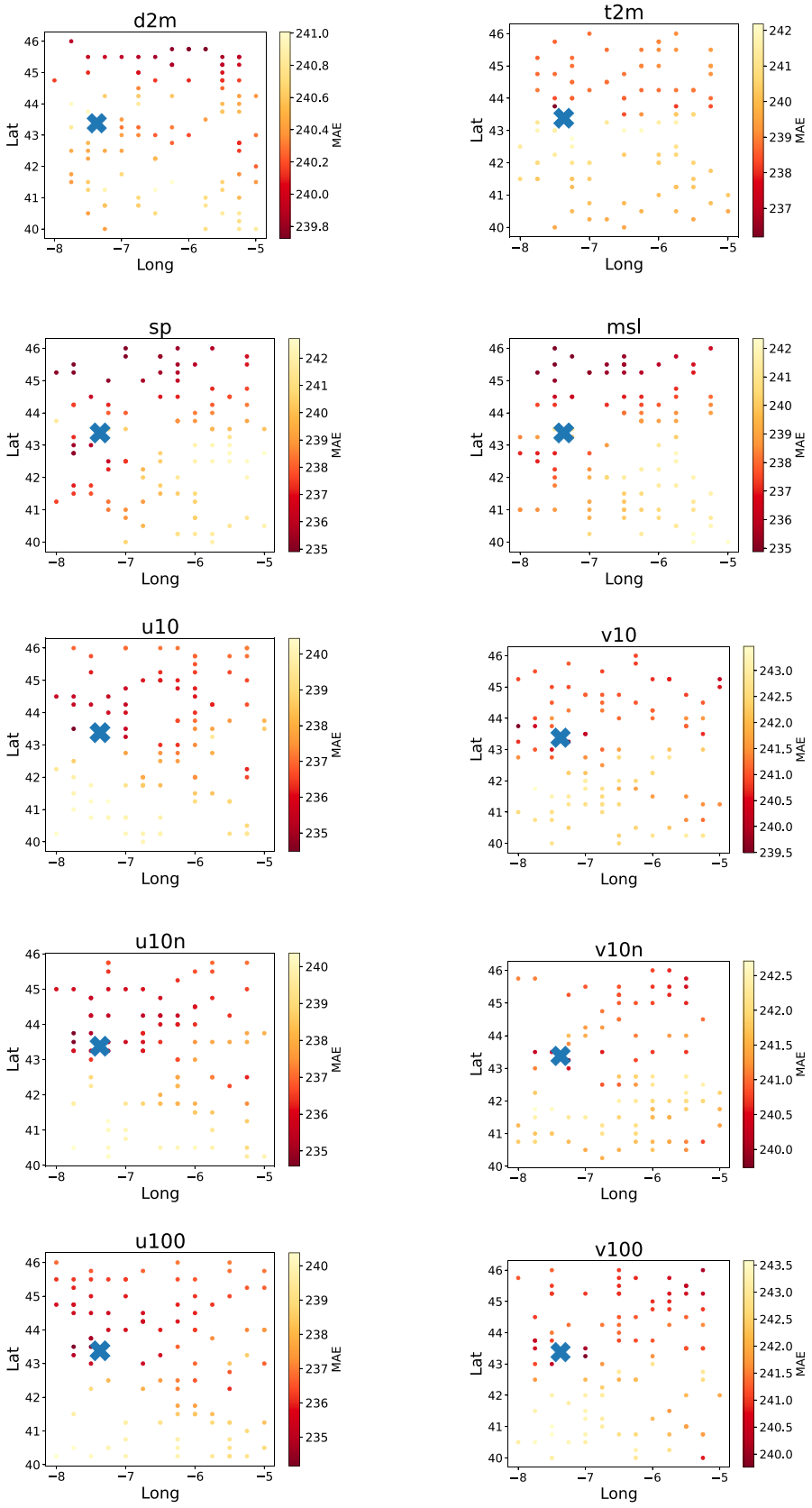

Fig. 9. Sensitivity analysis of geographical location of reanalisis nodes for each meteorological variable and using LR as regression method.

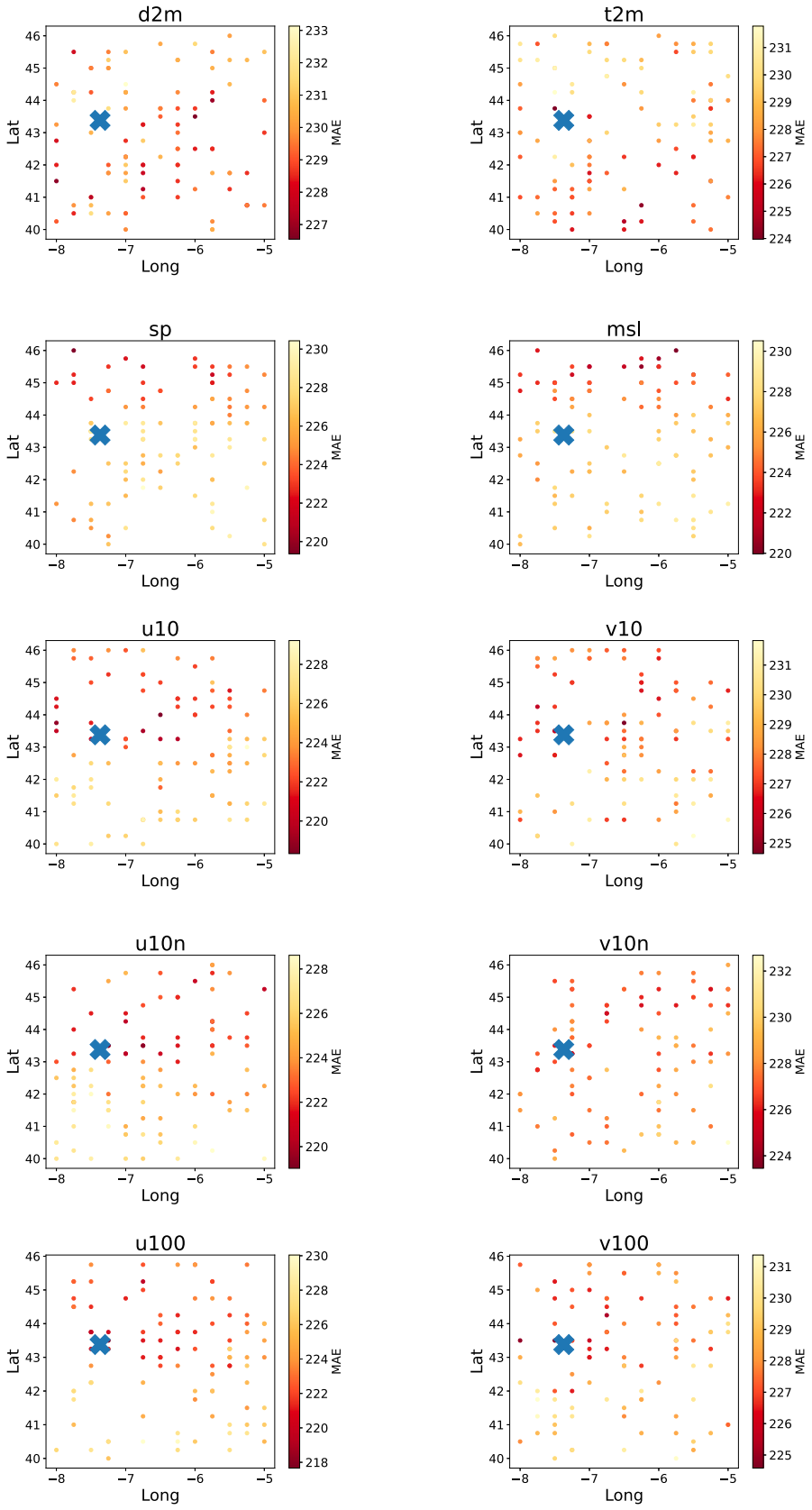

Fig. 10. Sensitivity analysis of geographical location of reanalysis nodes for each meteorological variable and using ELM as regression method.

Table 5
Codification of the solution provided by the optimization algorithms.

	Node 1	Node 2	Node 3	···	Node N
Variable 1	1	$m+1$	$2{\cdot}m+1$	···	$N{\cdot}m+1$
Variable 2	2	$m+2$	$2{\cdot}m+2$	···	$N{\cdot}m+2$
Variable 3	3	$m+3$	$2{\cdot}m+3$	···	$N{\cdot}m+3$
⋮	⋮	⋮	⋮	···	⋮
Variable m	m	$2{\cdot}m$	$2{\cdot}m+m$	···	$N{\cdot}m+m$

explains why some negative values appear in the improvement percentage. Besides, it is clearly noticeable the great difference between applying RS as the optimization algorithm versus applying DE or PSO, yielding significantly superior results in the three time horizons (with PSO performing slightly better in all cases). The better performance of these evolutionary algorithms is also evidenced when we observe the number of variables incorporated in the models, where a huge difference is appreciated between RS versus DE and PSO. Finally, it may be noted that, for all the regression methods, the performance improvement with the addition of reanalysis variables increases as the prediction time horizon is extended, indicating that in these cases the prediction

Table 6
Results for the five regression methods and the three optimization algorithms considered for a prediction time-horizon of 1 h.

	RS			DE			PSO		
	MAE	(%)	NºVar	MAE	(%)	NºVar	MAE	(%)	NºVar
LR	234.67	**1.62**	3.60	229.66	**3.72**	214.40	228.00	**4.42**	195.00
RT	233.90	−1.72	2.80	229.17	0.34	13.40	230.67	−0.31	11.00
RF	194.98	−0.99	2.80	194.36	−0.67	11.35	191.76	0.68	5.25
ELM	220.17	0.56	3.00	219.09	1.05	2.40	221.35	0.03	2.20
FCDNN	**160.42**	0.61	1.30	**159.98**	0.89	1.40	**160.36**	0.65	1.20
Average	208.83	0.02	2.70	206.45	1.07	48.59	206.43	1.09	42.93

Table 7
Results for the five regression methods and the three optimization algorithms considered for a prediction time-horizon of 3 h.

	RS			DE			PSO		
	MAE	(%)	NºVar	MAE	(%)	NºVar	MAE	(%)	NºVar
LR	458.24	2.46	2.80	418.78	**10.86**	235.20	414.82	**11.70**	318.30
RT	444.48	−0.18	3.60	441.55	0.48	11.50	441.51	0.49	15.00
RF	414.87	−1.25	2.70	390.84	4.62	5.00	386.70	5.63	10.20
ELM	409.22	**2.81**	2.60	396.99	5.71	5.10	395.01	6.18	7.50
FCDNN	**321.40**	0.38	4.00	**316.96**	1.76	2.00	**319.61**	0.94	2.00
Average	409.64	0.84	3.14	393.02	4.69	51.76	391.53	4.99	70.60

Table 8
Results for the five regression methods and the three optimization algorithms considered for a prediction time-horizon of 6 h.

	RS			DE			PSO		
	MAE	(%)	NºVar	MAE	(%)	NºVar	MAE	(%)	NºVar
LR	602.07	2.29	3.60	510.45	**17.15**	209.50	509.57	**17.30**	267.50
RT	596.15	1.74	2.70	585.11	3.56	6.50	580.88	4.25	11.70
RF	553.06	2.36	5.00	512.31	9.56	17.20	525.39	7.25	13.00
ELM	513.36	**9.51**	6.90	496.00	12.57	6.20	497.51	12.30	5.10
FCDNN	**437.33**	0.29	4.20	**430.11**	1.93	5.00	**402.07**	8.33	6.20
Average	540.39	3.24	4.48	506.80	8.95	48.88	503.08	9.89	60.70

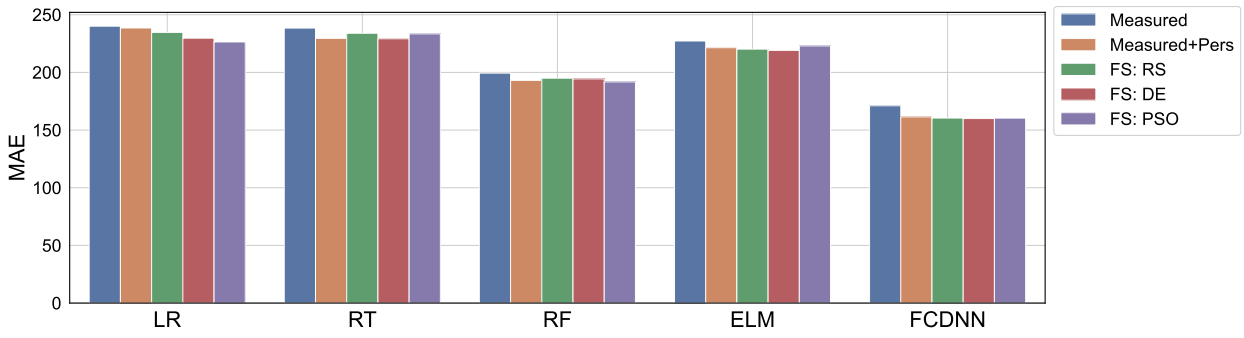

Fig. 11. Comparison of the models performance (MAE) using the different predictive variables considered, for a time forecast horizon of 1 h.

improvement margin is broader. Even so, it becomes evidenced that the use of reanalysis variables for fog prediction means a significant improvement for all cases.

Figs. 11–13 illustrate the comparison of the performance between all the models studied. We can first observe that for the three time horizons, FCDNN clearly outperforms the rest of regressors. Furthermore, it is noticeable that for all regression methods the addition of reanalysis variables leads to a significant improvement in the performance of the models, noting how this improvement becomes more significant as the prediction time horizon is shifted away.

Lastly, a temporal comparison between the measured visibility values and those predicted by the FCDNN regressor after being trained with the addition of reanalysis meteorological variables is shown in Figs. 14–16 for a time horizon of the prediction of 1, 3 and 6 h, respectively. It is possible to observe in these figures how the prediction degrades as more long-term prediction horizons are used. With an excellent prediction provided in the first case, a remarkable prediction in the second case, detecting the great majority of low visibility events, although with a lower accuracy in the regression; and a prediction with room for improvement for the last case, detecting some events with high performance (from hour 90 to 120), but also forecasting a high rate of false positives (from hour 390 to 500).

4. Discussion

In this section, a discussion on the iterative feature selection algorithm results is carried out, analyzing the meteorological variables chosen by the evolutionary algorithms and the importance of the geographical location of these Reanalysis variables.

The iterative forward selection algorithm was applied using two evolutionary optimization algorithms (DE and PSO) and employing 5 different regression methods (LR, RT, RF, ELM and FCDNN). This process was repeated 5 times for each case and with 3 different prediction time horizons. For each of these horizons, the meteorological variables selected by the evolutionary algorithms over the 50 runs of the feature selection algorithm (2 *optimization algorithms* × 5 *regressors* × 5 *executions*) were stored. The analysis of these variables is provided subsequently.

First, the 10 reanalysis meteorological variables considered in this study (Table 2), were divided into 4 groups: those related to temperature variables (t2m and d2m), to pressure (sp and msl), to wind at 10 meters altitude (u10, u10n, v10 and v10n) and to wind at 100 meters altitude (u100 and v100). Then, the distribution of each group of variables was determined. Figs. 17–19 displays the percentage of each group of variables over the total number of variables selected by the evolutionary algorithms for each time forecast horizon.

In all cases, two groups of variables are found to be much more meaningful than the rest. This indicates that when it comes to forecast fog events, wind variables at 10 meters are the most significant meteorological variables (from the ones considered) for the regression methods, followed by temperature, wind at 100 meters and, lastly, pressure.

Subsequently, an analysis of the selected Reanalysis nodes location was conducted. Similarly as in the previous case, for all 50 runs of the forward selection algorithm executed for each time horizon, the locations of the selected reanalysis variables were stored. Figs. 20–22 show the density plots of these selected nodes, in the form of contour plots, for the three prediction time-horizons. In addition, histograms relative to the longitude and latitude of these selected nodes are shown at the top and right of the figures, respectively. Reminding the location of the Mondoñedo meteorological station (43.38°N, 7.37°W), it may be observed how for 1 h time horizon, the location of the most significant nodes for the prediction are found at the south and west of Mondoñedo, being at this locations were most of the selected nodes are distributed. For the second prediction time-horizon (3 h), the most relevant nodes are displaced slightly to the southwest of Mondoñedo, while for the last case (6 h) they move a bit to the north towards the west of Mondoñedo.

These figures reveal that there is a relationship in the location of the reanalysis nodes and the performance of the regression models, being more significant the nodes placed in the south and in the west of the location where the visibility is intended to be forecasted.

5. Conclusions

In this paper the prediction of low-visibility events due to orographic fog at A-8 motor road (Galicia, Northern Spain), has been carried out using machine and deep learning algorithms, for three different prediction time-horizons: 1, 3 and 6 h). First, the forecast has been performed using meteorological variables measured in situ at the target weather station, where it was found that the DL models evaluated (FCDNN) significantly outperformed the other ML methods tested. The influence of adding previous hours visibility related variables has been evaluated, finding that, in this type of orographic fog, persistence plays a key role, with average improvements in prediction error up to 2.24%. This increase is especially noticeable for closer prediction horizon, achieving improvement rates of up to 5.65%.

Then, the influence of including additional exogenous meteorological variables from ERA5 reanalysis database among the predictors was analyzed, in order to improve the accuracy of the forecast. Preliminary work in this sense has consisted of a sensitivity analysis, where the dependence between the performance of the regression methods and the meteorological variables and geographical reanalysis nodes has been established. We have concluded that, in fact, there exist specific nodes location and meteorological variables that performs better that others. This reveals the need of applying a feature selection algorithm, for selecting the optimal meteorological variables and reanalysis nodes, according to each regression method and each time horizon. For this purpose, and due to the complexity of the task, a forward selection algorithm has been proposed and analyzed in this work. When applying this method with an optimization algorithm based on random search, and two evolutionary algorithms, Differential Evolution (DE) and Particle Swarm Optimization (PSO), a large improvement is achieved by

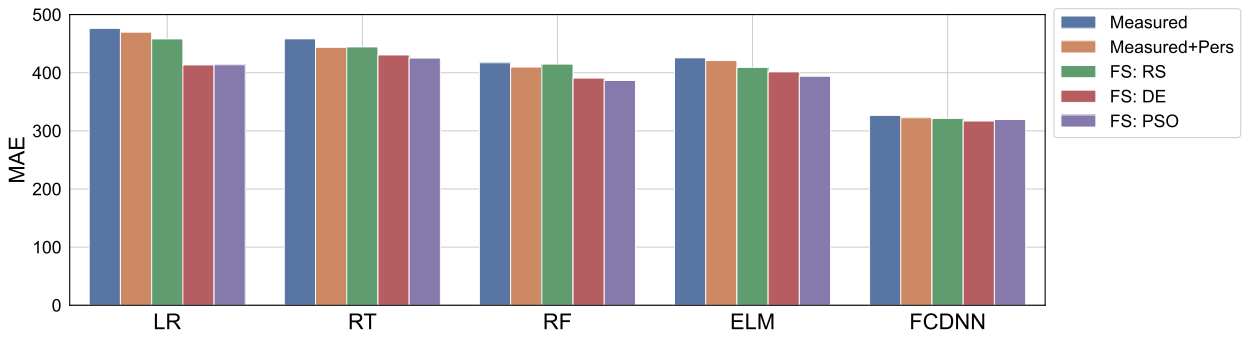

Fig. 12. Comparison of the models performance (MAE) using the different predictive variables considered, for a time forecast horizon of 3 h.

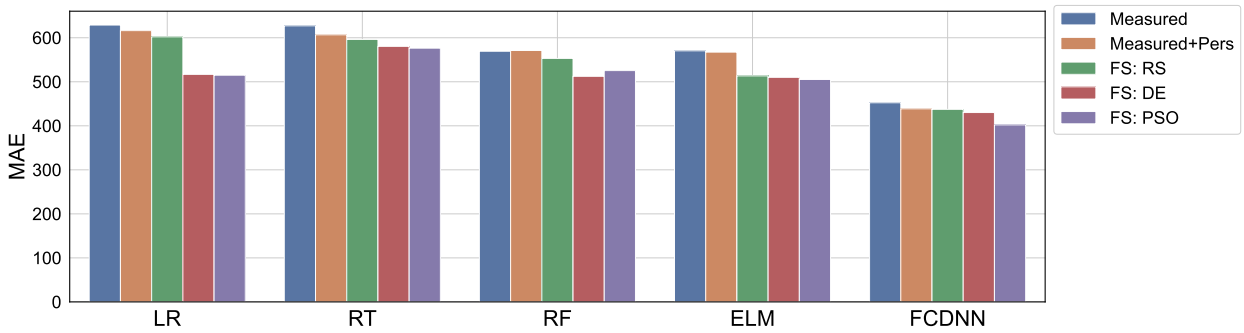

Fig. 13. Comparison of the models performance (MAE) using the different predictive variables considered, for a time forecast horizon of 6 h.

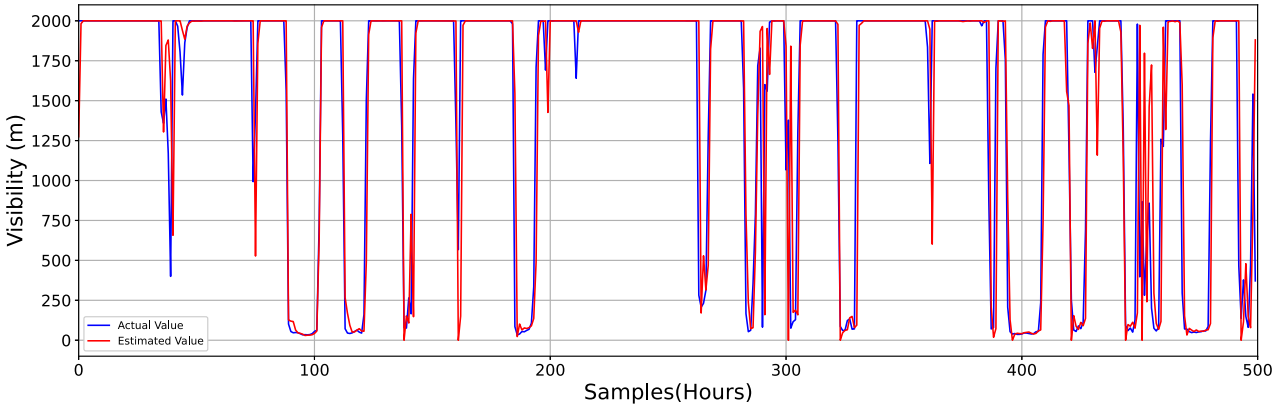

Fig. 14. Temporal comparison of actual measured visibility and predicted with a prediction time-horizon of 1 h using FCDNN as regressor, and including the selected reanalysis variables after applying the feature selection algorithm with PSO as optimizer.

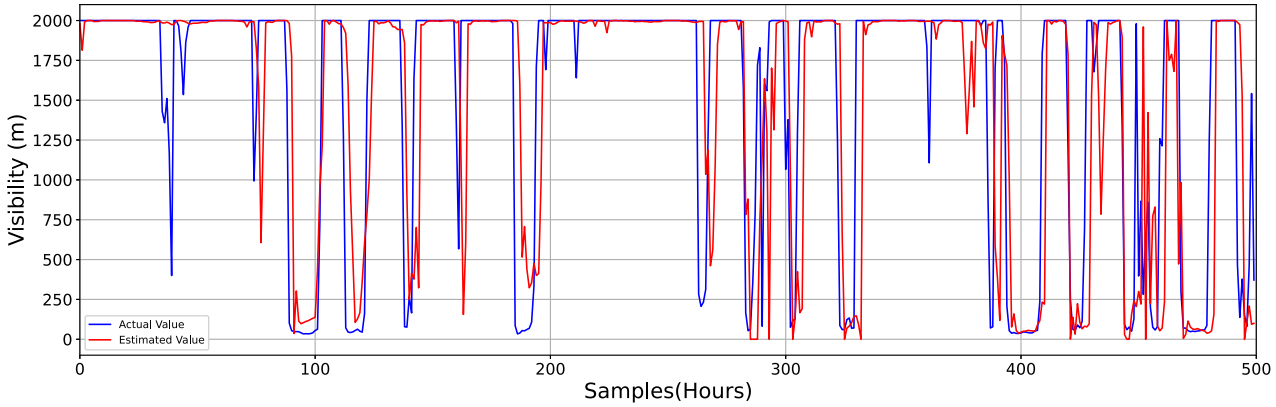

Fig. 15. Temporal comparison of actual measured visibility and predicted with a prediction time-horizon of 3 h using FCDNN as regressor, and including the selected reanalysis variables after applying the feature selection algorithm with PSO as optimizer.

using the reanalysis variables. This improvement is even more pronounced as the time horizon increases, reaching an average performance of 1.09% of improvement respect to the initial method in terms of MAE for the PSO usage and a 1-h horizon, a 4.99% for the 3-h horizon and a 9.89% for the 6-h horizon. In addition, it is shown that the evolutionary methods are very useful, since the number of variables selected is about 16 times higher than by using Random Search (RS). When comparing the performance of the two optimization algorithms assessed, it may be concluded that the number of variables selected by them is very similar in all cases, although the PSO achieves slightly better forecast results in all cases.

Finally, it is analyzed the type of variables chosen by the forward selection methodology and their location. Thus, it can be observed that the most representative variables (of those considered) are the 10 meters wind and the air temperature, which are the variables with the highest percentage of selection by the algorithm. Regarding the geographical location of the nodes selected, it has been observed that the most significant regions are located to the south and west of the target weather station, where the visibility is being predicted.

Thus, in this paper we have shown that the inclusion of exogenous meteorological variables, extracted from reanalysis, significantly improves the prediction of low visibility events, being a key aspect the selection of optimal variables and their location depending on the regression method and the time horizon used.

Future research lines will consist of, firstly, increasing the search space size for the evolutionary algorithms, adding a larger number of

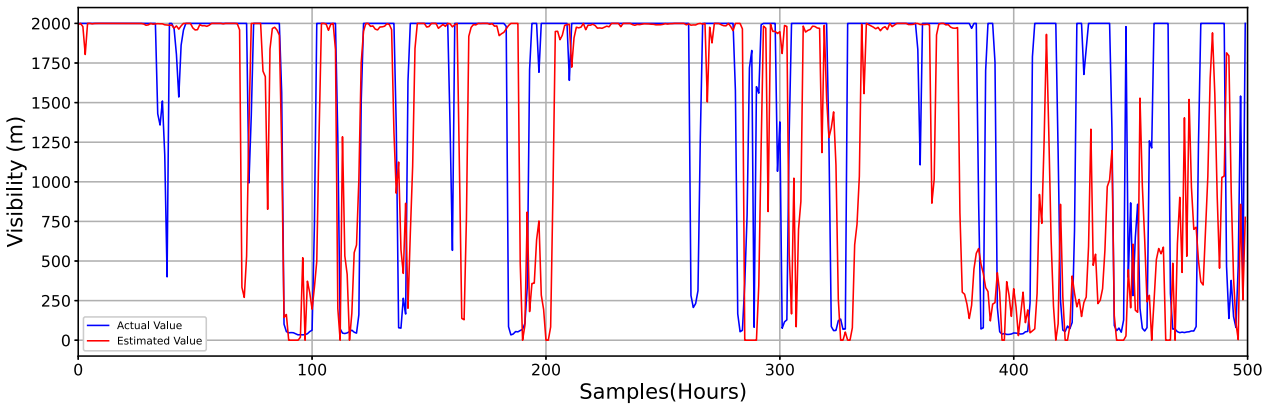

Fig. 16. Temporal comparison of actual measured visibility and predicted with a prediction time-horizon of 6 h using FCDNN as regressor, and including the selected reanalysis variables after applying the feature selection algorithm with PSO as optimizer.

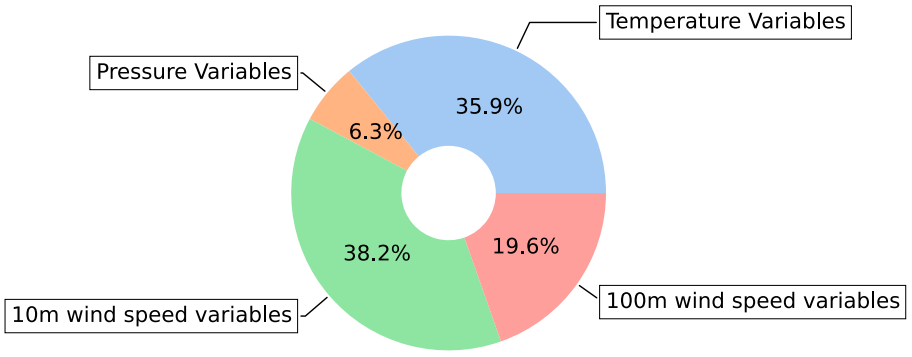

Fig. 17. Percentage of each group of variables over the total number of variables selected by the iterative forward selection algorithm for a time horizon of 1 h.

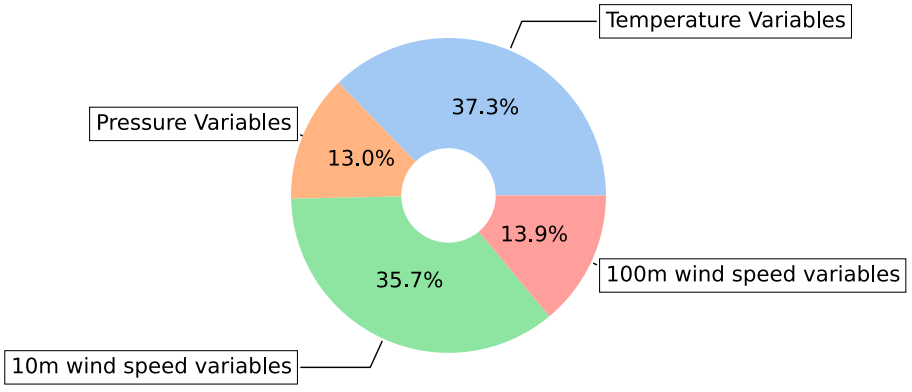

Fig. 18. Percentage of each group of variables over the total number of variables selected by the iterative forward selection algorithm for a time horizon of 3 h.

meteorological variables, and larger geographic regions. Also, the possibility for the algorithms to search in the past time series of each variable, so that to predict the fog at $T = h + 1$, the algorithm would be able to compare the performance of a variable at $T = h$ or at $T = h - n$. In addition, the application of this feature selection algorithm with deep learning methods for multivariate time series regression will the following step of this work.

CRediT authorship contribution statement

C. Peláez-Rodríguez: Software, Methodology, Conceptualization, Investigation, Writing-original-draft. **J. Pérez-Aracil:** Conceptualization, Software, Validation, Writing-review-editing. **C. Casanova-Mateo:** Writing-review-editing, Investigation, Methodology. **S. Salcedo-Sanz:** Writing-review-editing, Supervision, Project-administration, Funding-acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

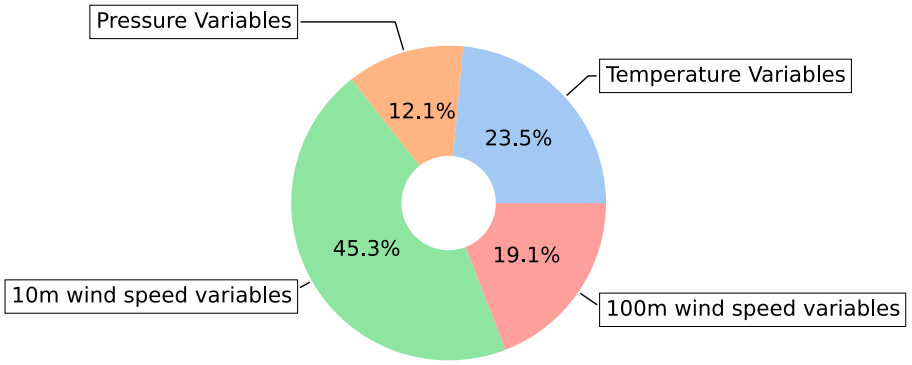

Fig. 19. Percentage of each group of variables over the total number of variables selected by the iterative forward selection algorithm for a time horizon of 6 h.

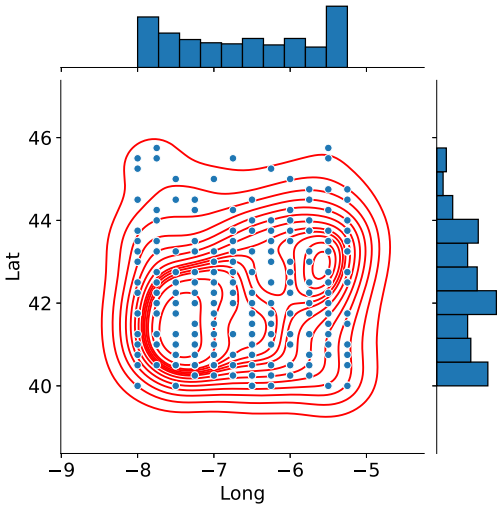

Fig. 20. Contour plot of the selected reanalysis nodes by the iterative forward selection algorithm for a time horizon of 1 h. Histograms relative to the longitude and latitude of these selected nodes are shown at the top and right of the figure, respectively.

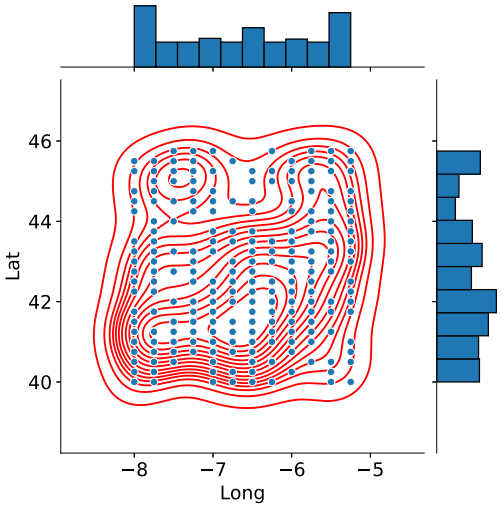

Fig. 21. Contour plot of the selected reanalysis nodes by the iterative forward selection algorithm for a time horizon of 3 h. Histograms relative to the longitude and latitude of these selected nodes are shown at the top and right of the figure, respectively.

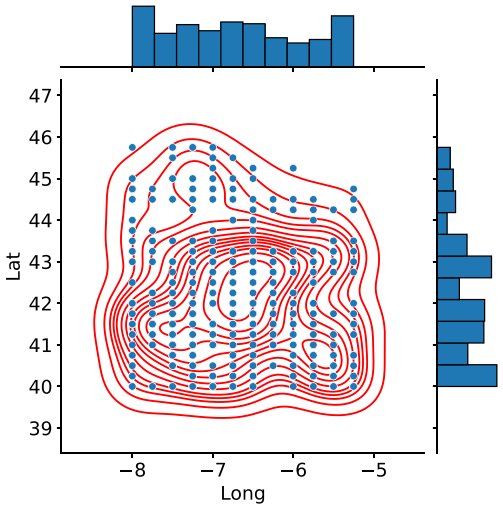

Fig. 22. Contour plot of the selected reanalysis nodes by the iterative forward selection algorithm for a time horizon of 6 h. Histograms relative to the longitude and latitude of these selected nodes are shown at the top and right of the figure, respectively.

Data availability

Data will be made available on request.

Acknowledgements

This research has been partially supported by the project PID2020-115454GB-C21 of the Spanish Ministry of Science and Innovation (MICINN). The authors would like to thank GSJ SOLUTIONS, part of SANJOSE Group, for providing the A8 weather and visibility data.

References

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A., Asari, V.K., 2019. A state-of-the-art survey on deep learning theory and architectures. Electronics 8 (3), 292.

Baldocchi, D., Waller, E., 2014. Winter fog is decreasing in the fruit growing region of the central valley of california. Geophys. Res. Lett. 41 (9), 3251–3256.

Bari, D., Ouagabi, A., 2020. Machine-learning regression applied to diagnose horizontal visibility from mesoscale NWP model forecasts. SN Appl. Sci. 2 (4), 1–13.

Bergot, T., Guedalia, D., 1994. Numerical forecasting of radiation fog. Part I: Numerical model and sensitivity tests. Mon. Weather Rev. 122 (6), 1218–1230.

Boneh, T., Weymouth, G.T., Newham, P., Potts, R., Bally, J., Nicholson, A.E., Korb, K.B., 2015. Fog forecasting for melbourne airport using a bayesian decision network. Weather Forecast. 30 (5), 1218–1233.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutierrez, P., Deo, R., Salcedo-Sanz, S., 2022. Machine learning regression and classification methods for fog events prediction. Atmos. Res. 272, 106157.
Cho, H.-J., Kim, K.-S., 2005. Development of hazardous road fog index and its application. J. East. Asia Soc. Transp. Stud. 6, 3357–3371.
Choi, W., Park, J., Kim, D., Park, J., Kim, S., Lee, H., 2022. Development of two-dimensional visibility estimation model using machine learning: Preliminary results for south korea. Atmosphere 13 (8), 1233.
Cornejo-Bueno, S., Casillas-Pérez, D., Cornejo-Bueno, L., Chidean, M.I., Caamaño, A.J., Cerro-Prada, E., Casanova-Mateo, C., Salcedo-Sanz, S., 2021. Statistical analysis and machine learning prediction of fog-caused low-visibility events at a-8 motor-road in spain. Atmosphere 12 (6), 679.
Dietz, S.J., Kneringer, P., Mayr, G.J., Zeileis, A., 2019. Forecasting low-visibility procedure states with tree-based statistical methods. Pure Appl. Geophys. 176 (6), 2631–2644.
Dismuke, C., Lindrooth, R., 2006. Ordinary least squares. Methods Des. Outcomes Res. 93 (1), 93–104.
Draper, N.R., Smith, H., 1998. Applied regression analysis, vol. 326. John Wiley & Sons.
Durán-Rosal, A.M., Fernández, J.C., Casanova-Mateo, C., Sanz-Justo, J., Salcedo-Sanz, S., Hervás-Martínez, C., 2018. Efficient fog prediction with multi-objective evolutionary neural networks. Appl. Soft Comput. 70, 347–358.
Fabbian, D., De Dear, R., Lellyett, S., 2007. Application of artificial neural network forecasts to predict fog at canberra international airport. Weather Forecast. 22 (2), 372–381.
Fernández-González, S., Bolgiani, P., Fernández-Villares, J., González, P., García-Gil, A., Suárez, J.C., Merino, A., 2019. Forecasting of poor visibility episodes in the vicinity of tenerife norte airport. Atmos. Res. 223, 49–59.
Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.
Ghimire, S., Deo, R.C., Downs, N.J., Raj, N., 2018. Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed modis satellite and reanalysis atmospheric products in solar-rich cities. Remote Sens. Environ. 212, 176–198.
Guerreiro, P.M., Soares, P.M., Cardoso, R.M., Ramos, A.M., 2020. An analysis of fog in the mainland portuguese international airports. Atmosphere 11 (11), 1239.
Gultepe, I., Tardif, R., Michaelides, S.C., Cermak, J., Bott, A., Bendix, J., Müller, M.D., Pagowski, M., Hansen, B., Ellrod, G., et al., 2007. Fog research: A review of past achievements and future perspectives. Pure Appl. Geophys. 164 (6), 1121–1159.
Hansen, B., 2007. A fuzzy logic–based analog forecasting system for ceiling and visibility. Weather Forecast. 22 (6), 1319–1330.
Herman, G.R., Schumacher, R.S., 2016. Using reforecasts to improve forecasting of fog and visibility for aviation. Weather Forecast. 31 (2), 467–482.
Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al., 1979. Era5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) 10.
Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70, 489–501. https://doi.org/10.1016/j.neucom.2005.12.126.
Huang, G.-B., Wang, D.H., Lan, Y., 2011. Extreme learning machines: a survey. Int. J. Mach. Learn. Cybern. 2 (2), 107–122.
Jeong, Y.-S., Shin, K.S., Jeong, M.K., 2015. An evolutionary algorithm with the partial sequential forward floating search mutation for large-scale feature selection problems. J. Oper. Res. Soc. 66 (4), 529–538.
Jonnalagadda, J., Hashemi, M., 2020. Forecasting atmospheric visibility using auto regressive recurrent neural network. In: 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, pp. 209–215.
Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks, vol. 4. IEEE, pp. 1942–1948.
Kim, J., Kim, S.H., Seo, H.W., Wang, Y.V., Lee, Y.G., 2022. Meteorological characteristics of fog events in korean smart cities and machine learning based visibility estimation. Atmos. Res. 275, 106239.
Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev.: Data Min. Knowl. Disc. 1 (1), 14–23.
Marzban, C., Leyton, S., Colman, B., 2007. Ceiling and visibility forecasts via neural networks. Weather Forecast. 22 (3), 466–479.
Ortega, L., Otero, L.D., Otero, C., 2019. Application of machine learning algorithms for visibility classification. In: 2019 IEEE International Systems Conference (SysCon). IEEE, pp. 1–5.
Ortega, L.C., Otero, L.D., Solomon, M., Otero, C.E., Fabregas, A., 2022. Deep learning models for visibility forecasting using climatological data. Int. J. Forecast.
Palvanov, A., Cho, Y.I., 2019. Visnet: Deep convolutional neural networks for forecasting atmospheric visibility. Sensors 19 (6), 1343.
Peláez-Rodríguez, C., Marina, C.M., Pérez-Aracil, J., Casanova-Mateo, C., Salcedo-Sanz, S., 2023. Extreme low-visibility events prediction based on inductive and evolutionary decision rules: An explicability-based approach. Atmosphere 14 (3), 542.
Peláez-Rodríguez, C., Pérez-Aracil, J., de Lopez-Diz, A., Casanova-Mateo, C., Fister, D., Jiménez-Fernández, S., Salcedo-Sanz, S., 2023. Deep learning ensembles for accurate fog-related low-visibility events forecasting. Neurocomputing 549, 126435.
Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization. Swarm Intell. 1 (1), 33–57.
Price, K., Storn, R.M., Lampinen, J.A., 2006. Differential evolution: a practical approach to global optimization. Springer Science & Business Media.
Román-Cascón, C., Steeneveld, G., Yagüe, C., Sastre, M., Arrillaga, J., Maqueda, G., 2016. Forecasting radiation fog at climatologically contrasting sites: evaluation of statistical methods and wrf. Q. J. R. Meteorol. Soc. 142 (695), 1048–1063.
Saavedra-Moreno, B., Salcedo-Sanz, S., Carro-Calvo, L., Gascón-Moreno, J., Jiménez-Fernández, S., Prieto, L., 2013. Very fast training neural-computation techniques for real measure-correlate-predict wind operations in wind farms. J. Wind Eng. Ind. Aerodyn. 116, 49–60.
Salcedo-Sanz, S., Pastor-Sánchez, A., Prieto, L., Blanco-Aguilera, A., García-Herrera, R., 2014. Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization–extreme learning machine approach. Energy Convers. Manage. 87, 10–18.
Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., García-Herrera, R., 2018. Feature selection in machine learning prediction systems for renewable energy applications. Renew. Sustain. Energy Rev. 90, 728–741.
Salcedo-Sanz, S., Deo, R.C., Cornejo-Bueno, L., Camacho-Gómez, C., Ghimire, S., 2018. An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the sunshine state of australia. Appl. Energy 209, 79–94.
Salcedo-Sanz, S., Piles, M., Cuadra, L., Casanova-Mateo, C., Caamaño, A., Cerro-Prada, E., Camps-Valls, G., 2021. Long-term persistence, invariant time scales and on-off intermittency of fog events. Atmos. Res. 252, 105456.
Seo, J.-H., Lee, Y.H., Kim, Y.-H., 2014. Feature selection for very short-term heavy rainfall prediction using evolutionary computation. Adv. Meteorol. 2014.
Steeneveld, G., Ronda, R., Holtslag, A., 2015. The challenge of forecasting the onset and development of radiation fog using mesoscale atmospheric models. Bound.-Layer Meteorol. 154 (2), 265–289.
Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 11 (4), 341–359.
Stull, R., 2015. An algebra-based survey of atmospheric science. University of British Columbia, Columbia.
Torres-López, R., Casillas-Pérez, D., Pérez-Aracil, J., Cornejo-Bueno, L., Alexandre, E., Salcedo-Sanz, S., 2022. Analysis of machine learning approaches' performance in prediction problems with human activity patterns. Mathematics 10 (13), 2187.
Wen, W., Li, L., Chan, P., Liu, Y.-Y., Wei, M., 2023. Research on the usability of different machine learning methods in visibility forecasting. Atmósfera 37.
Wilson, A.M., Barros, A.P., 2017. Orographic land–atmosphere interactions and the diurnal cycle of low-level clouds and fog. J. Hydrometeorol. 18 (5), 1513–1533.
Yu, Z., Qu, Y., Wang, Y., Ma, J., Cao, Y., 2021. Application of machine-learning-based fusion model in visibility forecast: A case study of shanghai, china. Remote Sens. 13 (11), 2096.
Zang, Z., Bao, X., Li, Y., Qu, Y., Niu, D., Liu, N., Chen, X., 2023. A modified RNN-based deep learning method for prediction of atmospheric visibility. Remote Sens. 15 (3), 553.
Zhai, B., Wang, Y., Wu, B., 2023. An ensemble learning method for low visibility prediction on freeway using meteorological data. IET Intel. Transport Syst. 1–14.
Zhang, Y., Wang, Y., Zhu, Y., Yang, L., Ge, L., Luo, C., 2022. Visibility prediction based on machine learning algorithms. Atmosphere 13 (7), 1125.
Zhu, L., Zhu, G., Han, L., Wang, N., et al., 2017. The application of deep learning in airport visibility forecast. Atmos. Clim. Sci. 7 (03), 314.