Water Resources Management (2024) 38:2455–2470
https://doi.org/10.1007/s11269-024-03779-y

Real-Time Evaluation of the Uncertainty in Weather Forecasts Through Machine Learning-Based Models

Carmen Calvo-Olivera[1] · Ángel Manuel Guerrero-Higueras[2] · Jesús Lorenzana[1] · Eduardo García-Ortega[3]

Received: 28 March 2023 / Accepted: 31 January 2024 / Published online: 17 February 2024

Abstract

Meteorological events have always been of great interest because they have influenced everyday activities in critical areas, such as water resource management systems. Weather forecasts are solved with numerical weather prediction models. However, it sometimes leads to unsatisfactory performance due to the inappropriate setting of the initial state. Precipitation forecasting is essential for water resource management in semi-arid climate and seasonal rainfall areas such as the Ebro basin. This research aims to improve the estimation of the uncertainty associated with real-time precipitation predictions presenting a machine learning-based method to evaluate the uncertainty of a weather forecast obtained by the Weather Research and Forecasting model. We use a model trained with ground-truth data from the Confederación Hidrográfica del Ebro, and WRF forecast results to compute uncertainty. Experimental results show that Decision Tree-based ensemble methods get the lowest generalization error. Prediction models studied have above 90% accuracy, and root mean square error has similar results compared to those obtained with the ground truth data. Random Forest presents a difference of -0.001 concerning the 0.535 obtained with the ground truth data. Generally, using the ML-based model offers good results with robust performance over more traditional forms for uncertainty calculation and an effective alternative for real-time computation.

✉ Carmen Calvo-Olivera
mcalo@unileon.es

Ángel Manuel Guerrero-Higueras
am.guerrero@unileon.es

Jesús Lorenzana
jesus.lorenzana@scayle.es

Eduardo García-Ortega
eduardo.garcia@unileon.es

[1] Supercomputing Center Castile and León, SCAYLE, Campus de Vegazana s/n, León 24071, Castilla y Léon, Spain

[2] Robotics Group, Universidad León, Campus de Vegazana s/n, León 24071, Castilla y Léon, Spain

[3] Atmospheric Physics Group (GFA). Environmental Institute, Universidad León, Campus de Vegazana s/n, León 24071, Castilla y Léon, Spain

Keywords Precipitation · Machine learning · Forecast · Uncertainty · Decision Tree

Abbreviations

AB	Adaptative Boosting
ARW	Advanced Research WRF
AUC	Area Under the Curve
B	Bagging
BC	Boundary Conditions
CHEbro	Ebro Hydrographic Confederation
DT	Decission Tree
GAN	Generative Adversarial Network
HPC	High Performace Computing
IC	Initial Conditions
KPI	Key Performance Indicator
LDA	Linear Discriminant Analysis
LR	Logistic regression
ML	Machine Learning
MoEv	Model Evaluator
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NMM	Nonhydrostatic Mesoscale Model
NWP	Numerical Weather Prediction
QDA	Quadratic Discirminant Analysis
RES	Spanish Supercomputing Network
RF	Random Forest
RMSE	Root of the Mean Square Error
ROC	Receiver Operating Characteristic
SAIH	Automatic Hydrological Information System
SCAYLE	Supercomputación Castilla y León
SGD	Stochastic Gradient Descent
WRF	Weather Research and Forecasting

1 Introduction

Precipitation is one of the meteorological phenomena with the most significant impact on weather-dependent human activities (Torres-López et al. 2022), water resources (Anik et al. 2023) or agricultural water scarcity (Liu et al. 2023), among others. Weather forecasts, and specifically precipitation forecasts, are inherently uncertain. This uncertainty comes from the model physics and its initial and boundary conditions. So every weather forecast has some degree of uncertainty (Lorenz 1963). For some applications, only forecasts with an uncertainty estimate are valuable. Although it is computationally highly demanding, the best method for estimating the reliability of individual forecasts is to perform a set of numerical weather simulations (Scher and Messori 2018). AI, specifically Machine Learning (ML), can be utilized to manage the uncertainty arising from input data. Notably, ML demands significant computational power primarily during the training phase, offering a potential alternative for calculating the uncertainty in weather forecasts.

In this regard, diverse approaches have been implemented. For example, some works in the literature apply ML techniques to different fields related to forecasting, focusing on wind-related predictions. (Irrgang et al. 2020) tries to predict the uncertainty associated using a supervised learning approach with a recurrent neural network trained and tested with data from 2012 to 2017. In (Kosovic et al. 2020), they aim to measure the uncertainty of wind forecasts obtained through NWP models. Finally, (Wang et al. 2019) aims to obtain the uncertainty associated with the temperature, relative humidity and wind speed for each weather station they get data from using a dataset with 3-year forecasts from 10 weather stations in the Beijing region (China). Other works, such as (Hafeez et al. 2020; Bogner et al. 2019; Yang et al. 2020), focus on energy-related predictions.

Precipitation forecasting is essential and some works propose ML techniques to obtain precipitation predictions for specific locations, such as using convolutional neural networks to predict rainfall in a flood-causing area of Iran (Afshari Nia et al. 2023), to use UltraBoost, Stochastic Gradient Descending and Cost Sensitive Forest classifiers for flood prevention in Romania (Costache et al. 2022) or to forecast irrigation water requirements by evaluating ML models (Mokhtar et al. 2023). In (Parviz et al. 2023) introduced improved hybrid models combined by SVR and GMDH, used for representing the nonlinear component of the precipitation in two weather stations in humid and semi-arid climates in Iran.

Finally, a variety of ML methods have been successfully implemented to solve the problem of weather predictions, both for classification and regression problems, demonstrating its advantages in this area by reviewing state-of-the-art ML concepts, their applicability to meteorological data, and their relevant statistical properties (Schultz et al. 2021). 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 in Spain. Some authors consider whether it is possible to completely replace current numerical weather models and data assimilation systems with deep learning approaches (Schultz et al. 2021). Using methods based on deep learning with artificial convolutional neural networks that are trained on past weather forecasts can be another solution to indicate whether the predictability is different than usual (Scher and Messori 2018).

This work focuses on precipitation, a meteorological risk for society since severe rainfall causes flooding or ruins crops. Knowing precipitation at a specific location allows for preventing its effects. Precipitation forecasting is crucial but challenging in numerical weather models due to the involvement of multiple physical parameterizations, including longwave and shortwave radiation, convection, microphysics in mixed phases, turbulence, and planetary boundary layer processes (Tapiador et al. 2019). Therefore, getting an uncertainty index for precipitation forecasts may help decision-makers decide preventive actions. However, getting an uncertainty index in real-time is impossible since we require ground-truth data to compute the error in a forecast. Consequently, we aim to get an uncertainty index for WRF precipitation forecasts using an ML-based prediction model instead of ground-truth data to compute the forecast error.

We present an 11-year dataset from 2008 to 2018 with WRF forecasts and ground-truth precipitation data in the Ebro basin (Spain) available online under the name "Assessment of uncertainty in weather forecasts". We use it to train our ML model, which allows us to calculate the real-time uncertainty index associated with precipitation forecasts to meet the need for ground truth. We also compare the results obtained with ground truth data and our model.

The rest of the document is organized as follows: Section 2 describes the experiments carried out and the materials and methods used to evaluate our proposal. Then, results are shown in Section 3 and discussed in Section 4. Finally, Section 5 summarizes the conclusions and future lines of research.

2 Materials and Methods

As mentioned above, we aim to get an ML-based uncertainty index for WRF precipitation forecasts. The sections below depict the uncertainty index computation, the data gathering, the classification models' fitting to calculate the uncertainty in real time, and the evaluation method.

2.1 Computation of Uncertainty

Calculating uncertainty in weather predictions is crucial as it provides a quantitative measure of the reliability and accuracy of the forecasts, enabling decision-makers to assess potential risks and plan more effectively in response to varying climatic events.

Evaluating the error in the precipitation forecasts from a prediction model is possible by applying a cost function as the Root of the Mean Square Error (RMSE) (Wang et al. 2019). RMSE, a commonly used metric in regression tasks, measures the error between two datasets: the predicted values from a model and the actual ground-truth data. It is particularly effective in penalizing larger errors. RMSE is calculated as shown in Eq. (1). m is the number of cells in the grid where the study area discretizes, x_i depicts the precipitation value predicted by the WRF model for cell i, and y_i depicts the actual precipitation value in cell i.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \tag{1}$$

We require ground-truth data collected through the appropriate instrumentation to calculate precipitation error. Unfortunately, this fact makes it impossible to calculate the error in real-time, which prevents it from being used as an uncertainty value. Our research intends to replace the y_i values, corresponding to the actual precipitation value mentioned above, for those provided by a prediction model developed using supervised ML techniques from real-world data. So, our uncertainty index ($\mathcal{U}$) will be computed as shown in Eq. (2), where p_i depicts the predicted value in cell i of our ML-based model.

$$\mathcal{U} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - p_i)^2} \tag{2}$$

Therefore, one of our objectives is to get an ML-based model. Fitting a supervised ML-based model requires gathering a dataset. Furthermore, this dataset must contain features –aka predictors– and target variables –aka labels. Predictors come from postprocessing the weather forecasts generated by the WRF model. Labels –"rain" or "no rain" in our research– come from ground-truth data in the study area.

2.2 Study Area

The Ebro Valley, in the Northeast of Spain (see Fig. 1), is one of the regions in Europe with the highest number of summer convective storms that cause intense and heavy rain

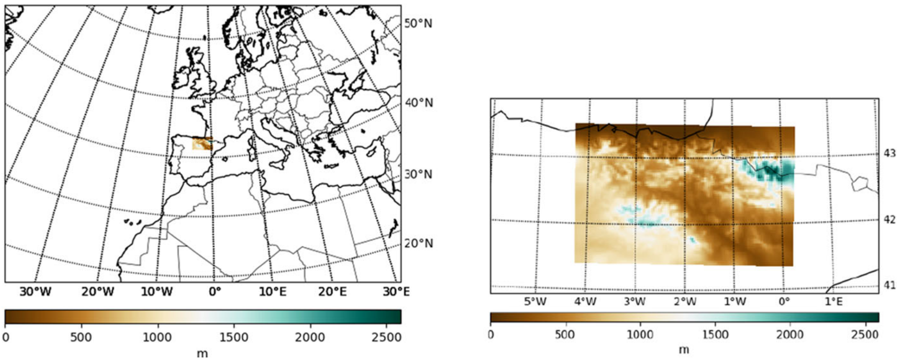


Fig. 1 Hydrographic Demarcation of the Ebro River Basin

and hail precipitation (García-Ortega et al. 2014). With an 80,000 km^2 area, it is the largest hydrographic basin in Spain. Besides, it presents a significant heterogeneity in its geology, topography and climate. The average annual rainfall varies from 2100 mm in the Pyrenees to 350 mm in the arid areas. The height varies from sea level on the Mediterranean coast to 3372 m at the Pyrenees. The topography varies from one basin to another (Samper et al. 2007).

The Automatic Hydrological Information System (SAIH) in the Ebro basin has a network of 367 meteorological stations collecting ground-truth data such as temperature (C) and rainfall (mm), among others. SAIH Ebro depends on the Ebro Hydrographic Confederation (CHEbro), which is the organism in charge of managing, regularising, and maintaining the waters and irrigations of the Ebro basin.

2.3 Data Gathering

To construct our ML-based model, it’s essential to compile a labeled dataset derived from the post-processing of WRF model weather forecasts. The dataset builds from the postprocessing of the weather forecasts generated by the WRF model. The WRF forecasts and their postprocessing have been carried out in the Caléndula, the supercomputer of Supercomputing Center Castile and León (SCAYLE), León (Spain) and one of the 17 supercomputers that conform to the Spanish Supercomputing Network (RES). The dataset [1] is available online.

2.3.1 Predictors

In the ML context, predictors are the input data mapped to a label through an empirical relationship. In our research, predictors correspond to meteorological variables computed by the WRF-ARW model (Skamarock et al. 2005). It carries out the complete workflow to assimilate observations into the model and runs in High-performance computing (HPC) environments.

The WRF model requires input data to set initial and boundary conditions (IC and BC). We gather input data from the National Centers for Environmental Prediction (NCEP) operational Global Forecast System (GFS) (National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce 2000). The NCEP operational

[1] https://doi.org/10.5281/zenodo.6421268

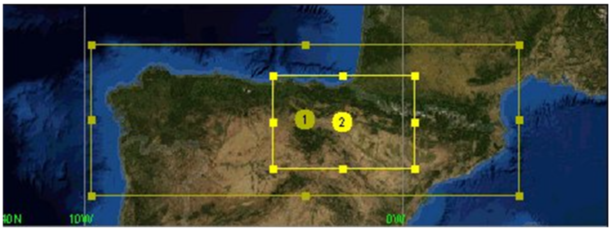


Fig. 2 Forecast domains

GFS analysis and forecast grids are on a 0.25 by 0.25 global latitude-longitude grid. Grids include analysis and forecast time steps at a three-hour interval from 0 to 240 and a 12-hour interval from 240 to 384.

After getting input data for IC and BC, we run the WRF model to get 24-hour forecasts from January 2008 to December 2018 with two nested domains of 9 × 9 and 3 × 3 km resolution, as shown in Fig. 2. There are different physics schemes available in the WRF model. Physics parameterization schemes describe sub-grid processes in numerical simulation models, such as the in-cloud microphysical processes responsible for precipitation (Tapiador et al. 2019). According to previous results in the study area (Merino et al. 2022), the Goddard Cumulus Ensemble one-moment bulk microphysics scheme (Tao et al. 1989, 2009) was selected. For cumulus, the Grell-Devenyi ensemble cumulus scheme (Grell and Dévényi 2002) for the outer domain. Convection in the inner domain was explicitly resolved. Other schemes selected were the Dudhia scheme (Dudhia 1989) for shortwave radiation, the Rapid Radiative Transfer Model (Mlawer et al. 1997) for longwave radiation, and the Noah Land Surface Model (Chen and Dudhia 2001).

The WRF outputs are in NetCDF format –intended to store multi-dimensional scientific data. Output files gather the values of up to 100 meteorological variables. We get the temperature and mixing ratio variables at different pressure levels (500 hPa, 700 hPa, and 850 hPa). The whole list is shown in Table 1. They will be our predictors. Besides, temperature and mixing ratio variables are not accumulative, so we get them with a 3-hour interval from 0 to 21 h. Since we consider seven variables at three pressure levels eight times daily, we get 168 values for each grid point in the study area depicted by its latitude, longitude, and height above the sea. We store our predictors in a separate NetCDF file for postprocessing once the labels are included within the dataset.

Table 1 Predictors

Predictor	Description
T	Temperature at 500 hPa, 700 hPa, and 850 hPa
QVAPOR	Column water vapour content at 500 hPa, 700 hPa, and 850 hPa
QCLOUD	Column liquid water content at 500 hPa, 700 hPa, and 850 hPa
QRAIN	Column rain at 500 hPa, 700 hPa, and 850 hPa
QICE	Column ice water vapour content at 500 hPa, 700 hPa, and 850 hPa
QSNOW	Column snow at 500 hPa, 700 hPa, and 850 hPa
QGRAUP	Column graupel at 500 hPa, 700 hPa, and 850 hPa

2.3.2 Labels

Once we gather our predictors, we must add the label –“rain” or “no rain”– to every sample in our dataset. We get such a label from the ground-truth data provided by the SAIH Ebro. The “rain” or “no rain” labels in our study are determined based on data collected by the rain gauges of the SAIH Ebro network, ensuring that our classifications accurately reflect the actual precipitation events recorded in the region. As mentioned, the SAIH Ebro manages a network of 367 meteorological stations that collect temperature (^oC) and rainfall (mm) values. Specifically, we gather rainfall values from 2008 to 2018 with their corresponding latitude and longitude coordinates. We apply the “rain” label when accumulated precipitation exceeds 0 mm and “no rain” otherwise.

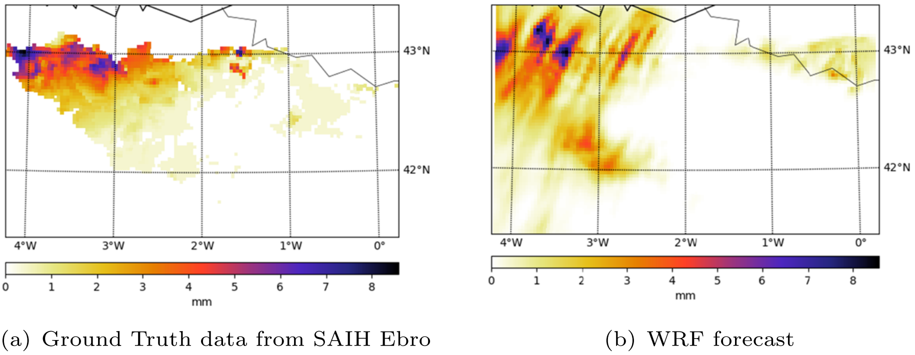
2.3.3 Post-Processing

Finally, some data curation is necessary to get the dataset. First, to set the suitable class for each sample, data must interpolate to the defined inner domain shown in Fig. 2. Therefore, we use the interpolation method described in (Merino et al. 2021). This approach involves rigorous quality control functions to identify and remove suspect data, interpolation techniques to align data within the defined study domain, and reconstruction methods for gap filling. This ensures that the dataset is not only comprehensive but also maintains high accuracy and reliability for subsequent analysis.

Next, after interpolating, the original dataset was filtered using quality control functions for identifying and removing suspect data. Lastly, reconstruction techniques for filling gaps (Serrano-Notivoli et al. 2017) prevent the lack of such data from affecting the experiments. For example, Fig. 3a shows the precipitation results on November 26th, 2008, from ground truth data from SAIH Ebro. Figure 3b shows the precipitation forecast by WRF for the same date.

2.4 Model Fitting

We aim to build a prediction model whose inputs are meteorological variables obtained from WRF forecasts and whose output p_i is a precipitation presence indicator for a specific grid point i in the study area.



(a) Ground Truth data from SAIH Ebro

(b) WRF forecast

Fig. 3 Precipitation estimate on November 26th, 2008

We use Model Evaluator (MoEv), a wrapper for the Scikit-Learn library (Pedregosa et al. 2011) to get our prediction models. MoEv has been successfully used in different research areas such as jamming attacks detection on real-time location systems (Guerrero-Higueras et al. 2018), or malicious-network-traffic detection (Campazas-Vega et al. 2020), among others.

We randomly split the dataset into 67% for the training set and 33% for the test set to get a training set for fitting the prediction models and a test set to ensure their generalization.

Then, we apply 10-fold cross-validation to fit prediction models. Since we need to predict a class –"rain" or "no rain", classification algorithms are more suitable than regression or clustering algorithms. However, since data matters more than algorithms for complex problems (Halevy et al. 2009) we aim to evaluate classification, clustering, and regression algorithms to select the most accurate for this problem.

Specifically, we compute: Adaptative Boosting (AB) (Freund and Schapire 1997), Decision Tree (DT) (Safavian and Landgrebe 1991), DT-based Bagging (DT-B) (Breiman 1996), Linear Discriminant Analysis (LDA) (Balakrishnama and Ganapathiraju 1998), Logistic Regression (LR) (Zhu et al. 1997), Quadratic Discriminant Analysis (QDA) (Hastie et al. 2009), Random Forest (RF) (Breiman 2001) and Stochastic Gradient Descent (SGD) (Bottou 2012). We selected these specific models based on their proven effectiveness in handling complex, non-linear relationships inherent in meteorological data. Models like Random Forests and Decision Trees are robust to outliers and capable of capturing intricate patterns in data. Despite its simplicity, Logistic Regression provides a strong baseline for performance comparison. The diversity of these models, ranging from ensemble methods to linear classifiers, allows for a comprehensive evaluation of different algorithmic approaches in accurately predicting precipitation events, ensuring the selection of the most effective model for our specific dataset and study objectives.

2.5 Evaluation

To evaluate our proposal, first, we need to assess the performance of our prediction models to get the most accurate. Therefore, we calculate well-known Key Performance Indicators (KPIs). First, models' performance is measured by considering their accuracy score as shown in Eq. 3), where T_P is the true-positive rate, T_N is the true-negative rate, F_P is the false-positive rate, and F_N is the false-negative rate.

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{3}$$

Besides, we consider the following KPIs obtained through the confusion matrix: Precision ($\mathcal{P}$), Recall ($\mathcal{R}$), and F$_1$-score ($\mathcal{F}_1$). $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ (Sokolova and Lapalme 2009) are computed as shown in Eqs. (4), (5), and (6). The $\mathcal{P}$ score shows the ratio between the number of correct predictions (both negative and positive) and the total number of predictions. The $\mathcal{R}$ score shows the rate of positive cases correctly identified by the algorithm. The $\mathcal{F}_1$ score relates to both $\mathcal{P}$ and $\mathcal{R}$ since it is their harmonic mean (Hossin and Sulaiman 2015).

$$\mathcal{P} = \frac{T_P}{T_P + F_P} \tag{4}$$

$$\mathcal{R} = \frac{T_P}{T_P + F_N} \tag{5}$$

$$\mathcal{F}_1 = 2\frac{\mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{6}$$

The chosen KPIs – Accuracy, $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}_1$ – are particularly relevant for our study as they provide a holistic assessment of the model's performance. Accuracy measures overall correctness, while precision and recall evaluate the model's ability to correctly predict rain events and avoid false alarms. We require a high $\mathcal{R}$ score, but only if the $\mathcal{P}$ score is also high enough to ensure that there are not too many false negatives. Thus, the F_1-score is crucial as it balances precision and recall, especially important in imbalanced datasets.

Moreover, to evaluate the tradeoff between the T_P and F_P rates, we compute Receiver Operating Characteristic (ROC) curve. Besides, We have calculated the Area Under the Curve (AUC). AUC assesses the model's ability to distinguish between the two classes ("rain" and "no rain"), ensuring that our model reliably predicts precipitation events, which is critical for effective weather forecasting.

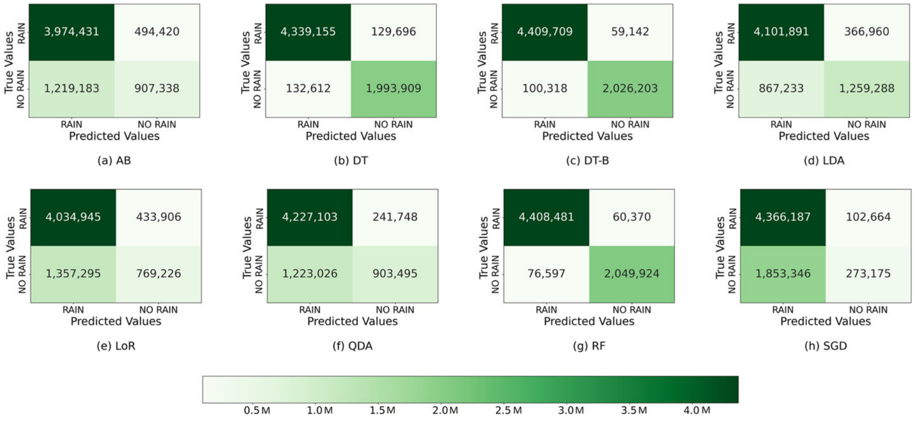
Finally, we can compute $\mathcal{U}$ according to Section 2.1 once we have selected the most accurate prediction model. Then, we carry out a statistical comparison between $\mathcal{U}$ and the actual RMSE for each WRF forecast from January 2008 to December 2018. Such analysis allows for measuring the performance of $\mathcal{U}$.

3 Results

Data gathering proposed in Section 2.3 allows for getting a 39 GB dataset of weather forecasts obtained by the WRF model from January 2008 to December 2018 in the study area. This dataset allows for fitting the prediction model we require to compute our uncertainty index $\mathcal{U}$ (see Section 2.1). The dataset contains 19,885,973 samples corresponding to a grid point – depicted by its latitude, longitude, and height – on a specific date. Each sample has the 168 features shown in Section 2.3.1 and the labels shown in Section 2.3.2. The dataset is available online.

Figure 4 displays the confusion matrices for the proposed prediction models, which are essential for computing the accuracy, precision, recall, and F_1-score values listed in Table 2. Besides, Fig. 5 shows the evaluated models' ROC curve and the AUC.

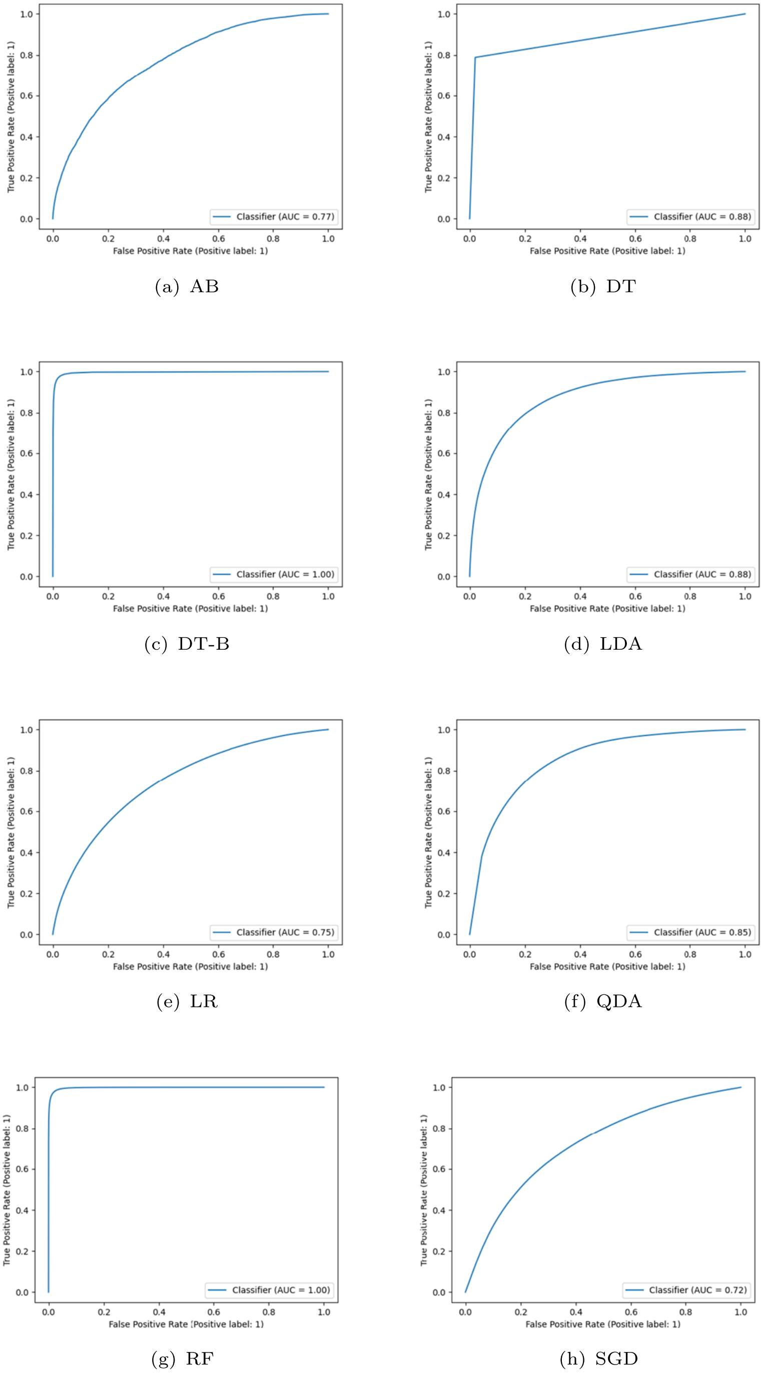
After computing the proposed KPIs, we can select the best prediction model to calculate our uncertainty index $\mathcal{U}$. Next, to evaluate $\mathcal{U}$, we compare it with the RMSE on every weather



(a) AB (b) DT (c) DT-B (d) LDA

(e) LoR (f) QDA (g) RF (h) SGD

Fig. 4 Confusion matrices of AB-, DT-, DT-B-, LDA-, LR-, QDA-, RF-, SGD-based prediction models



(a) AB

(b) DT

(c) DT-B

(d) LDA

(e) LR

(f) QDA

(g) RF

(h) SGD

Fig. 5 ROC and AUC

forecast from January 2008 to December 2018. Figure 6 presents the descriptive statistics obtained by computing the RMSE and $\mathcal{U}$ using different prediction models.

4 Discussion

The primary goal of this study was to establish the reliability of weather forecasts without the need for real-time data. This was achieved by focusing on three main aspects: first, developing an ML-based model as an alternative to ground truth data; then, creating a comprehensive and large dataset for training the model; and finally, comparing the results from our model with those obtained using ground truth data to assess our model's reliability.

To address the challenge of evaluating the uncertainty of weather forecasts obtained by the WRF model without real-time ground truth data, we sought to find an ML-based prediction model as a substitute. The reliability of this model was paramount, hence the computation of $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}_1$, ROC curves, and AUC to evaluate and select the most effective model. Previous studies have employed methods such as convolutional neural networks (Afshari Nia et al.

Table 2 Accuracy, precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F_1 score ($\mathcal{F}_1$) scores

Classifier	Accuracy	Class	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$
AB	0.740	No rain	0.765	0.889	0.823
		Rain	0.647	0.427	0.514
		Average	0.727	0.740	0.723
DT	0.960	No rain	0.970	0.971	0.970
		Rain	0.939	0.938	0.938
		Average	0.960	0.960	0.960
DT-B	0.976	No rain	0.978	0.987	0.982
		Rain	0.972	0.953	0.962
		Average	0.976	0.976	0.976
LDA	0.813	No rain	0.825	0.918	0.869
		Rain	0.774	0.592	0.671
		Average	0.809	0.813	0.805
LR	0.728	No rain	0.748	0.903	0.818
		Rain	0.639	0.362	0.462
		Average	0.713	0.728	0.703
QDA	0.778	No rain	0.776	0.945	0.852
		Rain	0.789	0.424	0.552
		Average	0.780	0.778	0.756
RF	0.979	No rain	0.983	0.986	0.985
		Rain	0.971	0.964	0.968
		Average	0.979	0.979	0.979
SGD	0.703	No rain	0.702	0.977	0.817
		Rain	0.727	0.128	0.218
		Average	0.710	0.703	0.624

2023), UltraBoost and Cost-Sensitive Forest classifiers (Costache et al. 2022), and tree-based algorithms (Yang et al. 2020), often applied in different work areas or with configurations unsuitable for WRF predictions.

In terms of data gathering, unlike other works (Wang et al. 2019; Ahmad et al. 2016) that made use of datasets not specific to our selected area or were limited in size, our approach involved creating a tailored dataset with predictors specifically chosen for our study area, as detailed in Table 1. Each sample in this dataset was labelled as “rain” or “no rain” using ground-truth data from SAIH Ebro, as explained in Section 2.1.

The confusion matrices for each model studied, represented in Fig. 4, show that the number of true positives is above 4 million for most models. Notably, the three classifiers with the highest accuracy, RF, DT-B, and DT, exhibit a similar rate of true negatives and low rates of failure (between 18% and less than 10%). This direct relationship between our study’s best and worst classifiers is evident in their performance in terms of false positives and negatives.

Table 2 presents the accuracy scores and the $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ scores for all tested classification models. While classifiers like LDA, LR, or SGD obtain satisfactory $\mathcal{R}$ scores, their $\mathcal{P}$ scores barely exceed 0.70. In contrast, models such as DT or DT-B show strong performance in both KPIs. The RF classifier emerges as the most effective, with the highest $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ scores, making it the best classifier according to our results.

The evaluated models’ ROC curves and AUC, as shown in Fig. 5, further confirm the superior performance of the RF and B classifiers. Both DT and LDA models exhibit an AUC of 0.88, demonstrating their effectiveness in class prediction. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), with TPR on the y-axis and FPR on the x-axis, providing a clear visual representation of model performance.

Finally, Fig. 6 displays a box-plot representation of the RMSE (Eq. 1) calculated for the ground truth and the $\mathcal{U}$ uncertainty index computed for the four best prediction models used. These models, with an accuracy higher than 80%, particularly DT, B, and RF, present plots and data akin to ground truth. LDA follows as the fourth model with the best accuracy rate. The results confirm that the models with the highest accuracy rates exhibit a $\mathcal{U}$ index similar to the RMSE of our ground truth, verifying the effectiveness of our approach.

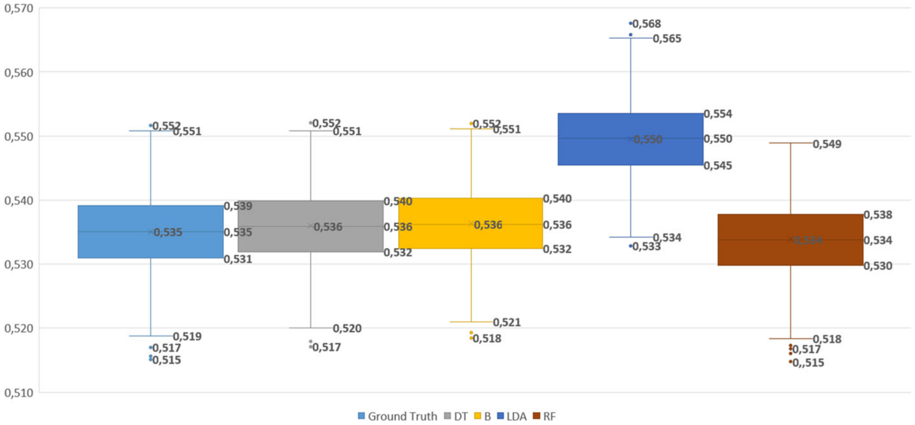


Fig. 6 RMSE for Ground Truth data and $\mathcal{U}$ with the best four prediction models

5 Conclusions

This study underscores the vital role of accurate weather forecasting in sectors like agriculture and water resource management. To enhance forecast precision, we utilized classification models and compiled an extensive dataset of precipitation data. The dataset, exceeding 40GB in CSV format and divided into two 20GB subsets, is publicly available under the Creative Commons Attribution 4.0 International license, providing a valuable resource for the scientific community.

Our methodology involved constructing a dataset from historical forecasts and ground truth data from the SAIH Ebro pluviometer network. This comprehensive dataset, encompassing precipitation data over 4,017 days and 9,594 points from domain 2 (refer to Fig. 2), was meticulously curated to remove missing values, resulting in 19,885,973 viable samples. We developed an RMSE-based index utilizing an ML-based prediction model instead of traditional ground-truth data to quantify forecast uncertainty. Various classification models were constructed using MoEv, a versatile wrapper for the Scikit-Learn library.

The experimentation demonstrated the efficacy of supervised learning algorithms in predicting the uncertainty of weather forecasts, fulfilling the primary goal of this research. Among the tested models, the RF classifier emerged as the most proficient in detecting precipitation, as evidenced by multiple KPIs. Renowned for its high generalization capability, RF's ensemble approach, integrating numerous decision trees with probability thresholds, proved suited for handling extensive data and many variables. Moreover, decision tree-based algorithms, namely DT and DT-B, exhibited superior performance compared to other evaluated models.

A notable breakthrough of this research is the assembly of a comprehensive dataset for the Ebro River basin region, spanning 11 years of WRF model forecasts. This dataset effectively supports the construction of models to estimate the uncertainty associated with a prediction. Additionally, we introduced a method to assess forecast uncertainty by calculating an uncertainty index $\mathcal{U}$. Our findings reveal that $\mathcal{U}$ closely aligns with values derived from ground truth, bolstering confidence in the forecast's reliability.

Looking beyond precipitation, future research might explore ML-based models' applicability to other meteorological variables, such as wind, hail, or snow, which significantly influence water resource variability and quality. Furthermore, contrasting our proposed models' performance and computational efficiency against neural network-based models could offer valuable insights into enhancing the reliability and speed of meteorological predictions.

Acknowledgements The research described in this article has been partially funded by Supercomputación Castilla y León (SCAYLE). The authors want to acknowledge the SAIH Ebro for providing temperature and rainfall data from its network of meteorological stations.

Author Contributions Conceptualization, C.C., Á.M.G, J.L. and E.G.; methodology, Á.M.G.; software, C.C.; validation, C.C. and Á.M.G.; formal analysis, Á.M.G.; investigation, C.C. and Á.M.G.; resources, J.L.; data curation, C.C. and J.L.; writing—original draft preparation, C.C. and Á.M.G.; writing—review and editing, J.L. and E.G.; visualization, C.C., Á.M.G, J.L. and E.G.; supervision, Á.M.G, J.L. and E.G.; project administration, Á.M.G, J.L. and E.G.; funding acquisition, Á.M.G. All authors have read and agreed to the published version of the manuscript.

Funding This work has been financially supported by the Ministry for Digital Transformation and of Civil Service of the Spanish Government through the QUANTUM ENIA project call - Quantum Spain project, and by the European Union through the Recovery, Transformation and Resilience Plan - NextGenerationEU within the framework of the Digital Spain 2026 Agenda.

Data Availability The datasets generated and/or analysed during the current study are available in the "Assessment of uncertainty in weather forecasts" repository, https://doi.org/10.5281/zenodo.6421268.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Afshari Nia M, Panahi F, Ehteram M (2023) Convolutional neural network-ann-e (tanh): A new deep learning model for predicting rainfall. Water Resour Manag 1–26

Ahmad A, Javaid N, Guizani M, Alrajeh N, Khan ZA (2016) An accurate and fast converging short-term load forecasting model for industrial applications in a smart grid. IEEE Trans Ind Inform 13(5):2587–2596

Anik AH, Sultan MB, Alam M, Parvin F, Ali MM, Tareq SM (2023) The impact of climate change on water resources and associated health risks in bangladesh: A review. Water Secur 18:100133

Balakrishnama S, Ganapathiraju A (1998) Linear discriminant analysis-a brief tutorial. Inst Signal Inf Process 18(1998):1–8

Bogner K, Pappenberger F, Zappa M (2019) Machine learning techniques for predicting the energy consumption/production and its uncertainties driven by meteorological observations and forecasts. Sustainability 11(12):3328

Bottou L (2012) Stochastic gradient descent tricks, Neural networks: Tricks of the trade, 421–436. Springer

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

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

Campazas-Vega A, Crespo-Martínez IS, Guerrero-Higueras ÁM, Fernández-Llamas C (2020) Flow-data gathering using netflow sensors for fitting malicious-traffic detection models. Sensors 20(24):7294

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

Chen F, Dudhia J (2001) Coupling an advanced land surface-hydrology model with the penn state-ncar mm5 modeling system. part i: Model implementation and sensitivity. Mon Weather Rev 129(4):569–585

Costache R, Arabameri A, Costache I, Crăciun A, Pham BT (2022) New machine learning ensemble for flood susceptibility estimation. Water Resour Manag 36(12):4765–4783

Dudhia J (1989) Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. J Atmos Sci 46(20):3077–3107

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

García-Ortega E, Hermida L, Hierro R, Merino A, Gascón E, Fernández-González S, Sánchez J, López L (2014) Anomalies, trends and variability in atmospheric fields related to hailstorms in north-eastern spain. Int J Climatol 34(11):3251–3263

Grell GA, Dévényi D (2002) A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. Geophys Res Lett 29(14):38–1

Guerrero-Higueras ÁM, DeCastro-García N, Matellán V (2018) Detection of cyber-attacks to indoor real time localization systems for autonomous robots. Robot Auton Syst 99:75–83

Hafeez G, Alimgeer KS, Wadud Z, Shafiq Z, Ali Khan MU, Khan I, Khan FA, Derhab A (2020) A novel accurate and fast converging deep learning-based model for electrical energy consumption forecasting in a smart grid. Energies 13(9):2244

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Cited on: 33

Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5(2):1

Irrgang C, Saynisch-Wagner J, Thomas M (2020) Machine learning-based prediction of spatiotemporal uncertainties in global wind velocity reanalyses. J Adv Model Earth Syst 12(5):e2019MS001876

Kosovic B, Haupt SE, Adriaansen D, Alessandrini S, Wiener G, Delle Monache L, Liu Y, Linden S, Jensen T, Cheng W et al (2020) A comprehensive wind power forecasting system integrating artificial intelligence and numerical weather prediction. Energies 13(6):1372

Liu J, Fu Z, Liu W (2023) Impacts of precipitation variations on agricultural water scarcity under historical and future climate change. J Hydrol 617:128999

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141

Merino A, García-Ortega E, Navarro A, Fernández-González S, Tapiador FJ, Sánchez JL (2021) Evaluation of gridded rain-gauge-based precipitation datasets: Impact of station density, spatial resolution, altitude gradient and climate. Int J Climatol 41(5):3027–3043

Merino A, García-Ortega E, Navarro A, Sánchez JL, Tapiador FJ (2022) Wrf hourly evaluation for extreme precipitation events. Atmos Res 274:106215

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA (1997) Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. Journal of Geophysical Research: Atmospheres 102(D14):16663–16682

Mokhtar A, Al-Ansari N, El-Ssawy W, Graf R, Aghelpour P, He H, Hafez SM, Abuarab M (2023) Prediction of irrigation water requirements for green beans-based machine learning algorithm models in arid region. Water Resour Manag: 1–24

National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce (2000) Ncep fnl operational model global tropospheric analyses, continuing from july 1999

Parviz L, Rasouli K, Torabi Haghighi A (2023) Improving hybrid models for precipitation forecasting by combining nonlinear machine learning methods. Water Resour Manag: 1–23

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674

Samper J, Alvares D, Pisani B, García M (2007) Evaluación del efecto del cambio climático en los recursos hídricos en la cuenca hidrográfica del ebro con gis-balan. *Ponencia presentada en las Jornadas de la Zona NO Saturada del Suelo. Córdoba*

Scher S, Messori G (2018) Predicting weather forecast uncertainty with machine learning. Q J R Meteorol Soc 144(717):2830–2841

Schultz M, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen L, Mozaffari A, Stadtler S (2021) Can deep learning beat numerical weather prediction? Philos Trans R Soc A 379(2194):20200097

Serrano-Notivoli R, de Luis M, Beguería S (2017) An r package for daily precipitation climate series reconstruction. Environ Modell Softw 89:190–195

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, Powers JG (2005) A description of the advanced research wrf version 2. Technical report, National Center For Atmospheric Research Boulder Co Mesoscale and Microscale

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437

Tao WK, Anderson D, Chern J, Entin J, Hou A, Houser P, Kakar R, Lang S, Lau W, Peters-Lidard C et al (2009) The goddard multi-scale modeling system with unified physics. In: Annales Geophysicae, Volume 27, pp 3055–3064. Copernicus GmbH

Tao WK, Simpson J, McCumber M (1989) An ice-water saturation adjustment. Mon Weather Rev 117(1):231–235

Tapiador FJ, Roca R, Del Genio A, Dewitte B, Petersen W, Zhang F (2019) Is precipitation a good metric for model performance? Bull Am Meteorol Soc 100(2):223–233

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

Wang B, Lu J, Yan Z, Luo H, Li T, Zheng Y, Zhang G (2019) Deep uncertainty quantification: A machine learning approach for weather forecasting. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2087–2095

Yang F, Wanik DW, Cerrai D, Bhuiyan MAE, Anagnostou EN (2020) Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration. Sustainability 12(4):1525

Zhu C, Byrd RH, Lu P, Nocedal J (1997) Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS) 23(4):550–560