Université Paris Cité

BIG DATA PROJECT

Weather Forecasting using Neural Network

Master VMI / DCI

27/04/2023

Fait par:
BOUSSA Maria
SADADOU Mohamed

2022/2023

ABSTRACT — Weather forecasting refers to the prediction of future atmospheric conditions based on current and historical data. This process involves the use of complex computer models that take into account various factors such as temperature, humidity, wind speed and direction, atmospheric pressure, and precipitation patterns. The applications of weather forecasting are numerous and varied. Some of the most common uses of weather forecasting include agriculture: farmers use weather forecasts to plan when to plant crops, when to harvest, and how to irrigation ; transportation: weather forecasts help airlines, shipping companies, and other transportation industries plan their routes and schedules ; energy: weather forecasts help energy companies determine how much power to generate and when to do so.
Overall, weather forecasting has become an essential tool for a wide range of industries and applications.
The aim of this study is to evaluate various techniques sourced from multiple articles, with the goal of training a dataset to forecast a specific variable utilizing weather data.

1. Introduction

As we already said, weather forecasting has important practical applications in many fields, including agriculture, transportation and energy. The ability to accurately predict the weather can help people plan their activities, make informed decisions, and stay safe in hazardous conditions. Furthermore, advances in technology and data analysis have made it possible to improve the accuracy of weather forecasts, which presents exciting opportunities for further research and innovation in this field.
However, the problem with weather forecasting is the inherent unpredictability of the atmosphere, which makes accurate predictions a challenging task. Weather patterns are influenced by a vast number of variables, such as temperature, pressure, humidity, wind speed and direction, and topography. Additionally, the climate is constantly changing, which makes it difficult to rely on historical data for accurate forecasts. Moreover, it can be challenging to collect and analyze large amounts of data, and to develop computer models that can accurately simulate the atmosphere.

There are numerous studies on the subject. Many researchers are attempting to find solutions to problems that may arise during weather predictions by developing increasingly sophisticated models.

2. Related Work

Our study draws upon six primary articles that present weather forecasting methods utilized in diverse fields.

2.1 The Articles

a) Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations

This article is written by Christina Brester , Viivi Kallio-Myers , Anders V. Lindfors, Mikko Kolehmainen and Harri Niska.

It discusses the challenges and potential of solar energy, particularly its dependence on weather conditions and the importance of accurate forecasting for smart grid management and energy storage optimization. The article examines various approaches to forecasting, including physics-based, data-driven, and hybrid models, and highlights the use of Numerical Weather Prediction (NWP) variables as predictors in day-ahead solar PV forecasting. The article explores the limitations of NWP data availability and cost, leading to the development of "NWP free" approaches that rely on historical weather observations for forecasting. The article introduces three scenarios for day-ahead solar PV forecasting: the Baseline scenario that uses NWP data for training and testing models, the Alternative scenario that uses historical weather observations for training and NWP data for testing, and the Optimistic scenario that uses only historical weather observations for both training and testing. The article evaluates the accuracy of the forecasting models in each scenario, utilizing a fully connected feed-forward Artificial Neural Network (ANN) and benchmarking against physics-based solar PV forecasts. The study uses solar PV power data collected from three sites in eastern Finland and seven weather variables as predictors, and evaluates model performance using nested cross-validation.

b) Assessing urban heat island effects through local weather types in Lisbon's Metropolitan Area using big data from the Copernicus service

Written by Cláudia Reis , António Lopes and A. Santos Nouri, the article discusses the phenomenon of Urban Heat Islands (UHIs) caused by urbanization and impermeabilization, which can lead to heat "bubbles" in urban areas surrounded by cooler environments. As urban populations continue to grow, ensuring thermal comfort for urban inhabitants becomes a significant challenge, particularly for vulnerable groups. UHIs are driven

by local weather conditions, including wind speed/direction, cloud cover, and precipitation, which have been studied to understand their effects on UHI intensities and patterns. The article highlights the need to synthesize prevailing meteorological conditions to deepen the knowledge of UHI and its climatological science. The analysis of climate phenomena at meso and micro scales requires a fine mesh of meteorological data with a high temporal resolution, which is challenging to obtain. The article introduces Copernicus, a European Union Earth Observation Program that contains air temperature, specific humidity, relative humidity, and wind speed raster data for 100 European cities from 2008 to 2017, as a valuable dataset for UHI analysis in urban areas where field data is insufficient or nonexistent. The article focuses on Lisbon, a Mediterranean city where UHI has been studied, to analyze UHI in Lisbon's Metropolitan Area (LMA) by Local Weather Types (LWTs) using Copernicus data and a mesoscale meteorological network. The article aims to present a replicable methodology for detailed UHI analysis, analyze UHI in LMA with fine detail by LWTs, and validate UHI records generated from the Copernicus data using hourly data collected from a mesoscale meteorological network installed in Lisbon.

c) Benchmarking urban local weather with long-term monitoring compared with weather datasets from climate station and EnergyPlus weather (EPW) data

This article is written by Wei Wang , Shengguo Li , Siyi Guo , Min Ma , Shihu Feng and Li Bao.
It discusses the importance of accurately estimating urban building energy demand, especially as urbanization accelerates and energy consumption in cities increases. Building simulation technologies are commonly used to estimate energy consumption, but weather plays a significant role in building simulation accuracy. The article presents a comparison of three weather datasets (EnergyPlus Weather for TMY, Chinese Standard Weather Data, weather from urban climate stations, and weather from long-term measurement with a local micro-climate station) to uncover gaps and differences in weather datasets from different sources. The article highlights the importance of actual and local weather data for building energy analysis and suggests that researchers and building engineers choose suitable weather files for accurate building energy estimation.

d) An adaptive big data weather system for surface transportation

Written By Amanda R. Siems-Anderson, Curtis L. Walker , Gerry Wiener, William P. Mahoney III and Sue Ellen Haupt. This article is a research paper about Pikalert ®, which is an adaptive big data system that uses both historical and real-time observations for multiple, complex observing platforms including vehicles, and utilizes gridded forecasting systems to maintain safety in vehicle operation.

e) An innovative decision making method for air quality monitoring based on big data-assisted artificial intelligence technique

The authors, Leiming Fua, Junlong Lib and Yifei Chen, proposed a study that highlights the potential of Big Data and AI technologies in resolving severe environmental problems in China, specifically in air quality forecasting for environmental protection monitoring. The study proposes an innovative decision-making method for air quality monitoring by combining various algorithms in Big Data and AI, which addresses the limitations of a single AI algorithm in air quality forecasting. The research establishes two air quality forecasting models based on traditional Machine Learning methods and Deep Learning with atmospheric subject knowledge. The TSTM model based on Deep Learning shows better performance in all aspects than similar models, and the air forecast accuracy rate is higher even under heavy pollution. However, the study has limitations as it only monitored the air quality of some cities in Shaanxi Province, and future research will introduce the air quality data of the Beijing-Tianjin-Hebei region for verification. Overall, the study highlights the potential of Big Data and AI technologies in addressing environmental challenges and the importance of combining various algorithms to improve the accuracy of air quality forecasting models.

f) A decision support system for vessel speed decision in maritime logistics using weather archive big data

Written by Habin Leea, Nursen Aydin, Youngseok Choi, Saowanit Lekhavat, Zahir Irani.
The paper proposes a method for optimizing the speed of liner vessels to reduce fuel costs and greenhouse gas emissions, while maintaining service level agreements with cargo clients. The proposed method uses weather archive data to estimate real fuel consumption functions for speed optimization problems, and applies particle swarm optimization, a metaheuristic optimization method, to find Pareto optimal solutions that minimize fuel consumption and maximize service level. The paper also discusses the trade-off between vessel operation cost and service level for a given liner route, and outlines the steps involved in the proposed searching algorithm. The paper contributes to vessel speed optimization literature by proposing a novel method

to parse weather archive data and apply data mining techniques to learn the impact of weather conditions on fuel consumption.

2.2 Relation between the articles

Each of these articles is focused on weather forecasting, utilizing weather data and big data to address specific recurring issues within a given area. For instance, in Article **e)** [1], the authors explore the use of big data and artificial intelligence (AI) technology in environmental protection monitoring. Additionally, they propose a combined machine learning model for air quality forecasting to effectively tackle real-world challenges in environmental protection monitoring. Or in article **d)**[2], where the authors discuss the possibility of using vehicles to collect weather data. This opportunity has the potential to revolutionize the weather industry by significantly increasing the number of weather observations near the surface and offering unique datasets for deducing and extrapolating information about road conditions[3].

Moreover, the potential impact of weather and climate phenomena on various aspects of human life is an important theme that runs through all the texts. For example, accurate weather and climate prediction is important for optimizing energy production from renewable sources like solar energy, reducing energy consumption in buildings, and improving public safety in transportation. Climate phenomena like urban heat islands can have a significant impact on the health and comfort of urban inhabitants, and accurate climate data is important for environmental protection and monitoring. Therefore, understanding and predicting weather and climate phenomena is crucial for various aspects of human life and well-being.

While these articles explore a diverse range of applications for weather forecasting in various fields, a common theme is also the presence of data-related challenges. These difficulties can arise from issues with data quality and availability, which are particularly problematic given the significant quantities of data required by weather forecasting models. Additionally, certain meteorological phenomena, such as hurricanes or thunderstorms, can be incredibly intricate and tough to accurately model. Furthermore, our understanding of some atmospheric processes is still limited despite advancements in the field, which can complicate accurate forecasting. Weather conditions can also change quickly, and even small changes can have a notable impact on predictions. Finally, weather is inherently unpredictable, and even the most sophisticated models and data can only provide approximations of future conditions that are subject to change.

3. Problem Formulation

One of the main obstacles in modeling accurate PV forecasting is the availability of numerical weather prediction (NWP) data.
In the article “Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations”, the authors propose an alternative scenario where an artificial neural network (ANN) is trained on weather observations instead of NWP data and then tested on NWP data to simulate the model's use in operational PV forecasting. The authors conducted experiments with solar PV output data, historical weather observations, and historical NWP data collected from three sites in eastern Finland. The results show that, although training ANN on observational data leads to a slight decrease in performance compared to ANN trained on NWP data, it still outperforms a physical model. In practice, this means that if historical NWP data are not available for model training, observational data can still allow for effective model selection and parameter tuning, and generalization error estimates can be gradually updated using online NWP data.

The objective of the problem is to determine if the findings of the article are consistent with a broader range of data, by comparing the performance of three models: MLP, LSTM, and a combination of both. By evaluating the performance of the three models using a meteorological dataset covering Paris from 1945 to 2023, we compared their effectiveness under varying data processing approaches. We therefore decided to use this data set to try to predict the direct radiation measured in W/m^2 using all the other data set fields

We started with a basic pre-processing approach, having just scale the data, followed by the introduction of redundant data, voluntarily retaining all columns and adding additional columns to emphasize the desired relationships, and finally, implementing appropriate data processing. This comparison allowed us to observe how different data processing techniques influence the performance of the presented models when applied to the same data set.

[1] An innovative decision making method for air quality monitoring based on big data-assisted artificial intelligence technique

[2] An adaptive big data weather system for surface transportation

[3] https://ral.ucar.edu/solutions/products/pikalert

4. Materials and methods

Our solution primarily revolves around article one[4], therefore it is essential to provide a clear and comprehensive explanation of their actions.

4.1 Method provided in the first article

For recall, in this study, we are comparing the effectiveness of data-driven and physics-based methods for day-ahead solar PV forecasting in three sites located in eastern Finland. To achieve this, the researchers have devised three scenarios: the Baseline scenario, the Alternative scenario, and the Optimistic scenario. The Optimistic scenario uses only observational data for both training and testing. The objective of this study is to assess the accuracy of the forecasting models in each scenario through nested cross-validation.

The study employed a nested cross-validation approach to evaluate the performance of a fully connected feed-forward artificial neural network (ANN) in predicting solar PV power generation. The outer loop of cross-validation evaluated model performance in three different scenarios, while the inner loop selected model hyperparameters. The data were split into 5-fold intervals, with each fold containing data from all seasons.

We can break down their method into several steps, which are as follows:

1. They first constructed a fully connected neural network architecture and then trained and tested it.
2. The Adam optimization algorithm was applied to perform stochastic gradient descent. This algorithm uses the current averages and second moments of gradients to calculate adaptive learning rates for each parameter.
3. The Mean Absolute Error (MAE) was chosen as the loss function and computed using mini-batches of 128 instances during training.
4. To ensure consistency, they normalized the input data by transforming all variables to the range [0,1].
5. In the inner loop of nested cross-validation, they automatically determined the number of epochs, the number of neurons in the hidden layers, the number of previous timestamps included in the input vector, and the type of activation function in the hidden layers.
6. Finally, they passed the selected parameter set to the neural network model in the outer loop, where a multilayer perceptron (MLP) with a certain number of neurons in the hidden layers and the appropriate activation function will be trained on the training data with input variables obtained by concatenating predictors for a certain number of timestamps preceding the prediction point.

As we said earlier, data-driven predictive models were compared using three different scenarios: Baseline, Alternative, and Optimistic. The main difference between the scenarios was the source of the data used for training and testing the models. The Baseline scenario used NWP data for both training and testing, while the Alternative scenario used observational data for training and NWP data for testing. The Optimistic scenario used only observational data for both training and testing.
In all these scenarios, weather-related inputs of data-driven models were the same: temperature, global and diffuse radiation, wind speed and wind direction, precipitation intensity, and total cloud fraction[5].

For their data processing, they eliminated rows containing missing values, which could have otherwise skewed the results, and incorporated the time variable with day and time information. By examining their code and the resulting output, we were able to extract valuable insights from their dataset. We gathered their data to examine its structure. By scraping the HTML page where the API data was showcased, we retrieved each section in JSON format. Since their API encoded the data as float64, we converted it from float64 to float. Next, we used Python to save the results in CSV format and obtained the four Excel files.

4.2 Solution

To answer our problem, we decided to create and train various neural network architectures using TensorFlow and Keras for predicting direct radiation (solar radiation) from weather data.
MLP, LSTMc, and ArbitraryNN are all neural network architectures that can be used for weather forecasting.

- MLP is a type of feedforward neural network that can be used for time series forecasting by taking past values of weather variables as input and predicting future values of those variables. It can be

[4] Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations

[5]https://www.sciencedirect.com/science/article/pii/S0960148123002811#sec2

used for both regression and classification tasks.

- LSTMc is a type of recurrent neural network that can capture temporal dependencies in time series data. It is particularly effective when there are long-term dependencies between time steps, and can be used for both univariate and multivariate time series forecasting. LSTMs have been shown to be effective in weather forecasting, such as predicting rainfall, temperature, and wind speed.

- ArbitraryNN is a more flexible neural network architecture that can be customized to suit the specific characteristics of the weather data being forecasted. It can contain a combination of LSTM, Dense, and Dropout layers, and can be used for both univariate and multivariate time series forecasting. It can be especially useful when there are complex relationships between weather variables, or when the relationship between weather variables and their effects on other processes (such as plant growth or energy consumption) is not well understood.

The code has been divided into several sections, and we will explain each part.

1. An abstract class 'NeuralNet' is defined, which will be inherited by other specific neural network classes.
   The Neural net class is used to set the hyperparameters and variables in order to have them predefined during the initialization of the different models. Only the architecture of each model changes with the parameter [n;n] which specifies the number of layers and the number of neurons per layer.
2. We define MLP, LSTMc, and ArbitraryNN classes. These classes inherit from the abstract NeuralNet class and define their specific architecture by implementing the 'build_model' method.
3. The dataset is read from a CSV file, and various preprocessing steps are performed. Features are computed, such as mean_cloudcover, mean_windspeed, mean_pressure, mean_winddirection, mean_radiation, and a PCA component for precipitation and rain. We also scale the data. Columns not needed for analysis are dropped, and the dataset is split into features (X) and target (y) variables.
4. The data is split into training, validation, and testing sets using the train_test_split function from the sklearn library.
5. An instance of the MLP class is created with a specific architecture (32 neurons in each hidden layer and 'relu' activation). The model is then built and trained using the train_model method with 50 epochs.
6. The trained model is used to make predictions on the test set. A scatter plot is created to visualize the relationship between the true values and predicted values of direct radiation. The plot shows the measured values in red and the predicted values in blue.
7. Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Normalized Mean Absolute Error (nMAE), and Normalized Root Mean Squared Error (nRMSE) are computed to evaluate the model's performance.

5. Experiments and Results

In this project, we relied on a i5-8265U 16go ram DDR4 2400mhz 1tb HDD 256 SSD Intel UHD graphics AMD Radeon 520 graphics computer as our primary hardware. Jupyter and Google Colab are the main tools used to run the notebooks. We used Google Colab with GPU to accelerate code execution.

5.1 Dataset used

We obtained the dataset used in our analysis from Open-Meteo, an open-source weather API that offers free access for non-commercial purposes. The dataset contains various weather-related variables for the city of Paris, France, and covers the period from 1945 to 2023, with hourly resolution. Open-Meteo partners with national weather services to provide open data with a resolution ranging from 1 to 11 km. We then downloaded the .csv file containing the values for the features we selected.

We selected the following features as they represent various important weather-related variables that are commonly used in weather forecasting.:

Feature	Unity	Description
temperature_2m	°C	key variables that can impact human comfort and are important for a range of applications including agriculture, construction, and energy production.
relativehumidity_2m	%	

apparent_temperature	°C	takes into account the effect of wind on human perception of temperature, which is important in assessing comfort levels and for public health warnings.
pressure_msl	hPa	important for forecasting wind patterns and can provide information on the likelihood of storms, as well as changes in temperature and humidity.
surface_pressure	hPa	
precipitation	mm	provide information on the likelihood of precipitation and can help in forecasting weather patterns, such as the likelihood of thunderstorms or hailstorms.
rain	mm	
weathercode	Wmo code	
cloudcover	%	provide information on the amount and altitude of cloud cover, which can affect the amount of solar radiation that reaches the surface and thus impact temperature and other weather patterns.
cloudcover_low	%	
cloudcover_high	%	
shortwave_radiation	W/m²	provide information on solar radiation and can be useful in forecasting weather patterns related to temperature, humidity, and cloud cover.
direct_radiation (target)	W/m²	
diffuse_radiation	W/m²	
direct_normal_irradiance	W/m²	
windspeed_10m	km/h	provide information on wind patterns, which are important for a range of applications including agriculture, construction, and transportation.
windspeed_100m	km/h	
winddirection_10m	°	
winddirection_100m	°	

Overall, the selected features provide a comprehensive overview of various weather patterns and can help in developing accurate weather forecasting models.

5.2 Experiments

The authors evaluated the performance of their models using three types of data, including historical weather data, Numerical Weather Prediction (NWP) data, and a hybrid model. In an attempt to reproduce the authors' methodology, we managed to extract a portion of their data, using the method we explained in a previous section; however, the limited number of data points and their distribution made it challenging to conduct a conclusive experiment without incorporating k-fold validation. As the tests were performed directly and hyperparameters were manually adjusted, it wasn't feasible to utilize their dataset effectively for a comprehensive evaluation.

To run our experiments, we utilized the historical weather data that we described in the previous section.

The objective of these experiments is to emphasize and examine the effects of various data processing techniques on model performance when used with the same dataset. We accomplished this by assessing the prediction of the direct_radiation feature using the three methods previously discussed.

We started with a basic pre-processing approach (Experiment_1), followed by implementing proper data treatment (Experiment_2) and finally, introducing redundant data (Experiment_3). We then compared the results obtained for the three methods of MLP, LTSMc and ArbitraryNN.

B) Second experiment description

In our second experiment, we utilized our historical weather dataset and augmented it with additional features. The following steps were taken to carry out this experiment:

1. Compute the mean values for different groups of columns and adds new columns for the mean values:

- Mean of 'cloudcover (%)', 'cloudcover_low (%)', and 'cloudcover_high (%)' is computed and added as a new column 'mean_cloudcover'.
- Mean of 'windspeed_10m (km/h)' and 'windspeed_100m (km/h)' is computed and added as a new column 'mean_windspeed'.
- Mean of 'pressure_msl (hPa)' and 'surface_pressure (hPa)' is computed and added as a new column 'mean_pressure'.
- Mean of 'winddirection_10m (°)' and 'winddirection_100m (°)' is computed and added as a new column 'mean_winddirection'.
- Mean of 'shortwave_radiation (W/m²)', 'diffuse_radiation (W/m²)', and 'direct_normal_irradiance

(W/m²)' is computed and added as a new column 'mean_radiation'.

2. Perform PCA on 'precipitation (mm)' and 'rain (mm)' columns:
3. Scale the 'precipitation (mm)' and 'rain (mm)' columns using the StandardScaler.
4. Perform PCA with one principal component on the scaled data.
5. Add the principal component as a new column 'precipitation/rain PCA' in the DataFrame.
6. Drop the original 'precipitation (mm)' and 'rain (mm)' columns from the DataFrame.
7. Filter the DataFrame to keep only specific columns ('time', 'temperature_2m (°C)', 'relativehumidity_2m (%)', 'precipitation/rain PCA', 'weathercode (wmo code)', 'direct_radiation (W/m²)', and 'winddirection_100m (°)'), along with the new mean columns ('mean_cloudcover', 'mean_windspeed', 'mean_pressure', 'mean_winddirection', and 'mean_radiation').

After these transformations, the dataset will only contain the selected columns and the newly computed mean columns, as well as the 'precipitation/rain PCA' column, which is a single principal component representing the scaled 'precipitation (mm)' and 'rain (mm)' columns.

C) Third experiment description

In Experiment 3, we adopted a similar strategy for Experiment 1, but with the addition of redundant data. In summary, the main distinction between the two experiences is how we manage the precipitation columns and the radiation columns. In experiment 3, we have incorporated the mean radiation and the ACP of the precipitation, as well as the mean direction of the wind, while retaining the original columns, our goal is therefore to force the relations between the columns and to allow prediction, this creates a bias and distorts the results because the relationships are "easy" to find by the ANN.

5.3 Results

In order to assess the performance of our models, we employed the most pertinent evaluation metrics that we discovered in the literature we reviewed. Ultimately, we selected the following metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Normalized Mean Absolute Error (nMAE), and Normalized Root Mean Squared Error (nRMSE).

1. Mean Squared Error (MSE): It is the average of the squared differences between the predicted values and the actual values. This metric emphasizes larger errors, as squaring the differences amplifies the impact of large errors. It is commonly used in regression problems to measure the performance of a model. The lower the MSE, the better the model's performance.

2. Mean Absolute Error (MAE): It is the average of the absolute differences between the predicted values and the actual values. Unlike MSE, it does not amplify large errors, as it considers the absolute value of the differences. It is also used in regression problems to measure the performance of a model. The lower the MAE, the better the model's performance.

3. Root Mean Squared Error (RMSE): It is the square root of the Mean Squared Error (MSE). RMSE has the same unit as the predicted and actual values, making it easier to interpret than MSE. It also emphasizes larger errors, as it is derived from MSE. The lower the RMSE, the better the model's performance.
4. Normalized Mean Absolute Error (nMAE): It is the Mean Absolute Error (MAE) divided by the range of the actual values (difference between the maximum and minimum values). This metric allows for comparing the performance of models across different scales, as it expresses the error relative to the range of the actual values. The lower the nMAE, the better the model's performance.

5. Normalized Root Mean Squared Error (nRMSE): It is the Root Mean Squared Error (RMSE) divided by the range of the actual values. Similar to nMAE, it allows for comparing the performance of models across different scales by expressing the error relative to the range of the actual values. The lower the nRMSE, the better the model's performance.

We experimented with various hyperparameter settings and chose to present the most optimal outcomes we achieved.

5.3.1 Results for MLP

	Experiment 1	Experiment 2	Experiment 3
MSE	27.996	313.485	0.029
MAE	3.238	8.043	0.088
RMSE	5.291	17.705	0.170
nMAE	0.027	0.124	0.001
nRMSE	0.044	0.272	0.003

This table presents the results of three different experiments in terms of five evaluation metrics.

For experiment 1, MSE, MAE and RMSE suggest moderate errors in the model's predictions, nMAE and nRMSE indicate that the errors are relatively small compared to the range of the dataset, which suggests a good model performance.

And, for experiment 2, MSE, MAE and RMSE larger errors in the model's predictions compared to Experiment 1. While the other metrics show that the errors are larger compared to the range of the dataset, indicating that the model performance is worse than Experiment 1.

This comes from the fact that, by ignoring the precise columns and keeping only the mean and the PCA, the model must seek relationships and the lies are not obvious and easy to determine

For expirement 3, the metrics demonstrate an excellent model performance.

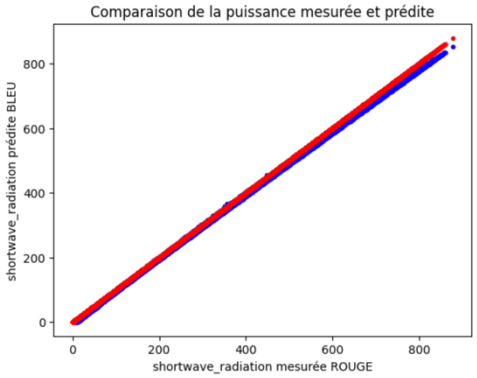

Figure 1: Predictions for MLP compared to the measured values for Experiment 1

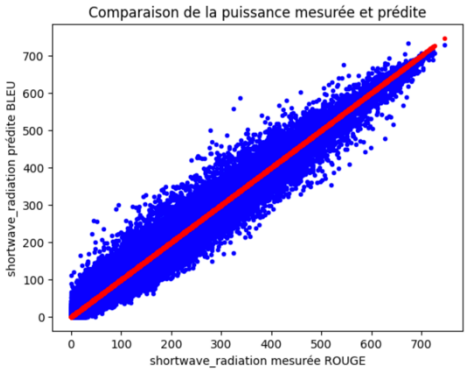

Figure 2: Predictions for MLP compared to the measured values for Experiment 2

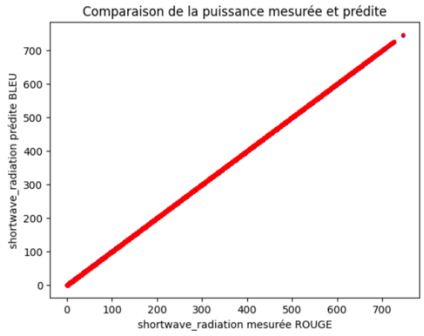

Figure 3 Predictions for MLP compared to the measured values for Experiment 3

In summary, the table and graphs show that Experiment 3 has the best performance of the model based on the five evaluation parameters, while Experiment 2 shows the worst performance. Experiment 1 is in between, with moderate errors and good performance when considering standard measurements.

However, these results are to be taken in perspective, in fact with the data provided for example to provide real precipitations as well as separed rain allows the model to determine the state of the rainy or snowy climate for example, out by providing only the average of the two it comes to say only if the climate is clear or not (clear therefore more brightness and therefore high irradiance) without determining with more precision and having an exact prediction

5.3.1 Results of LSTMc

	Experiment 1	Experiment 2	Experiment 3
MSE	297.610	315.108	4.966
MAE	9.721	8.074	1.226
RMSE	17.251	17.751	2.228
nMAE	0.081	0.124	0.019
nRMSE	0.144	0.273	0.034

Again, the table shows that Experiment 3 has the best performance of the model under the five evaluation parameters, while Experiment 2 shows the worst performance. Experiment 1 lies between the two, with significant errors, but better performance when considering the standardized measurements compared to the experiment,

These results may come from the fact that these types of models are adapted to the time series and thus find links with the data over a time period, out for the weather data that we have we can see that it is difficult to put forward this link because the quite changing weather from one period to another interferes, this suggests that hyperparametres without kfold validation are not sufficiently advanced in order to detect the complexity of the data, this confirms the conclusion of the article studied which suggests that this type of model for historical data remains limited compared to the MLP

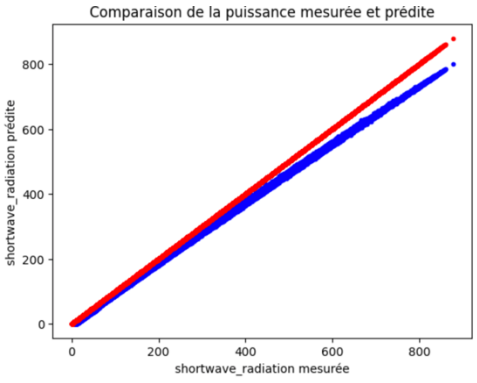

Figure 4 : Predictions for LSTMc compared to the measured values for Experiment 1

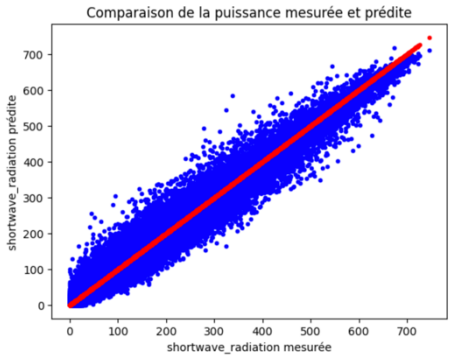

Figure 5: Predictions for LSTMc compared to the measured values for Experiment 2

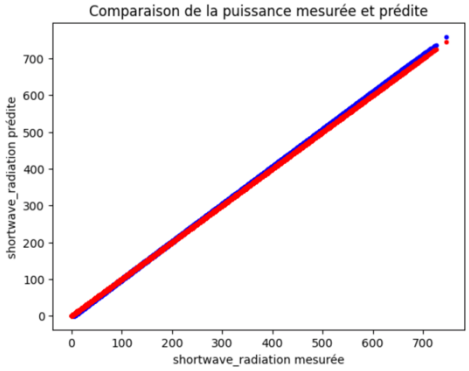

Figure 6 : Predictions for LSTMc compared to the measured values for Experiment 3

5.3.2 Results of Arbitrary NN

	Experiment 1	Experiment 2	Experiment 3
MSE	48170.458	367.871	380.056
MAE	119.606	9.537	9.604
RMSE	219.478	19.180	19.495
nMAE	0.999	0.147	0.148
nRMSE	1.833	0.295	0.300

This time, for experiment 1, MSE, MAE and RMSE suggest extremely important errors in the model predictions. While nMAE and nRMSE indicate that errors are almost equal to

The extent of the data set, which demonstrates very poor performance of the model.

this amounts to the model's low complexity compared to the data, with no Kfold validation and no more relationships between the data that the model fails to predict correctly, This summarizes the default values of the two models combined with LSTMc that do not achieve sufficient time series and MLP depth, resulting in inconclusive results.

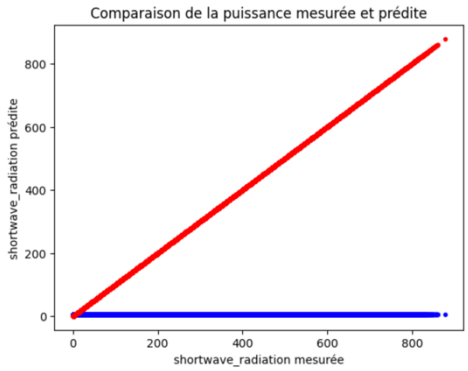

Figure 7 : Predictions for ArbitraryNN compared to the measured values for Experiment 1

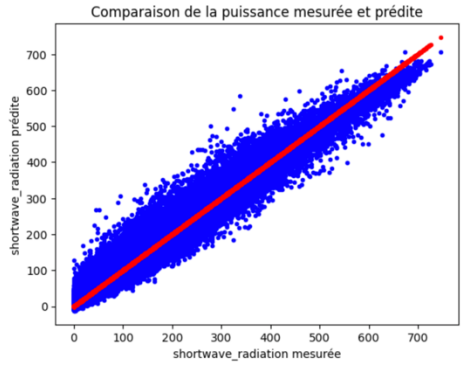

Figure 8: Predictions for ArbitraryNN compared to the measured values for Experiment 2

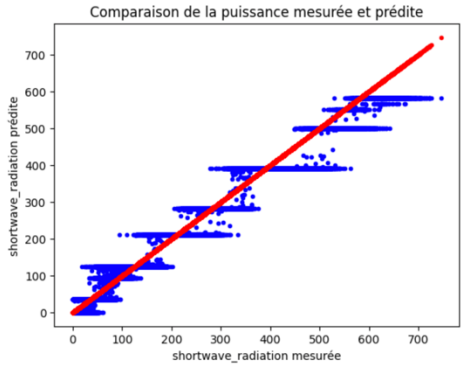

Figure 9 : Predictions for ArbitraryNN compared to the measured values for Experiment 3

The "staircase effect" visible in Figure 9 might be a result of saturation in the activation function. When this function gets saturated, the gradients become tiny during backpropagation, causing the learning process to slow down or even experience vanishing gradients. Consequently, the model may struggle to identify intricate patterns or pick up on subtle details in the data, leading to a stepped or staircase-like appearance in its predictions.

For the arbitrary NN model, saturation of its activation function could be a factor contributing to its lower performance compared to other models. this comes back to the defauts of the two models combined with the non-use of kfold validation and the two loops used in the article to adapt the hyperparameters

In the end, by incorporating the lessons learned from the graphs and their diagrams, we observe that the most realistic and close to real conditions model is experiment_3, which can be attributed to the specific approach we used for data processing. This is because there is no data redundancy that forces relationships, and in real-world scenarios it is more difficult to obtain separate and well-prepared data for different types of radiation. this is why it is more realistic to take into consideration processed data and without redundant columns to bias the results.

5.4 Global results

In conclusion, we observed improved performance for Experiment 2 and Experiment 3. This improvement can be attributed to the additional columns that were introduced to these datasets. However, it is important to note that these added columns might have introduced some bias, influencing the model's performance.

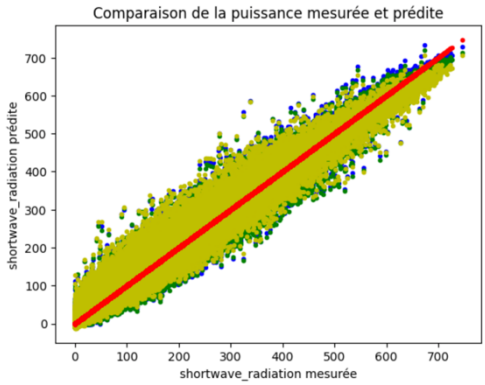

Figure 10 : Predictions of all methods tested compared to the measured values

6 Conclusion

In conclusion, weather forecasting plays a vital role in numerous fields, including agriculture, transportation, and energy. However, the complex nature of weather patterns, the inherent unpredictability of the atmosphere, and data-related challenges pose significant obstacles to accurate forecasting. Our study draws upon various articles to evaluate different techniques for training a dataset to forecast a specific variable using weather data. The results indicate that proper data treatment significantly impacts the performance of machine learning models in weather forecasting. The study also highlights the importance of using appropriate evaluation metrics to assess model performance accurately. Despite the challenges, weather forecasting remains an essential tool for various industries and applications, and continued research and innovation in this field are critical for improving the accuracy of predictions.

7 References

[1] Christina Brester, Viivi Kallio-Myers, Anders V. Lindfors, Mikko Kolehmainen, Harri Niska,

Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations. Link

[2] Cláudia Reis, António Lopes, A. Santos Nouri,

Assessing urban heat island effects through local weather types in Lisbon's Metropolitan Area using big data from the Copernicus service. Link

[3] Wei Wang, Shengguo Li, Siyi Guo, Min Ma, Shihu Feng, Li Bao,

Benchmarking urban local weather with long-term monitoring compared with weather datasets from climate station and EnergyPlus weather (EPW) data. Link

[4] Amanda R. Siems-Anderson, Curtis L. Walker, Gerry Wiener, William P. Mahoney, Sue Ellen Haupt,
An adaptive big data weather system for surface transportation. Link

[5] Leiming Fu, Junlong Li, Yifei Chen,
An innovative decision making method for air quality monitoring based on big data-assisted artificial intelligence technique. Link

[6] Habin Lee, Nursen Aydin, Youngseok Choi, Saowanit Lekhavat, Zahir Irani,

A decision support system for vessel speed decision in maritime logistics using weather archive big data. Link