environmental sciences proceedings

MDPI

Proceeding Paper

A Machine Learning Approach for Rainfall Nowcasting Using Numerical Model and Observational Data †

Georgios Kyros [1,2,*], Ioannis Manolas [1], Konstantinos Diamantaras [1], Stavros Dafis [2] and Konstantinos Lagouvardos [2]

[1] Department of Information and Electronic Engineering, International Hellenic University, 57400 Thessaloniki, Greece; manwlas.iwannis@gmail.com (I.M.); kdiamant@ihu.gr (K.D.)
[2] Institute of Environmental Research and Sustainable Development, National Observatory of Athens, 11810 Athens, Greece; sdafis@noa.gr (S.D.); lagouvar@noa.gr (K.L.)
* Correspondence: georgeky2001@gmail.com
† Presented at the 16th International Conference on Meteorology, Climatology and Atmospheric Physics—COMECAP 2023, Athens, Greece, 25–29 September 2023.

Abstract: The application of machine learning (ML) algorithms in large datasets in the field of meteorology is at the forefront of research. In this context, the use of satellite data to estimate the amount of rainfall is an important field of research, with operational applications. It is important to accurately predict the amount of rainfall (or rain rate) in a particular area for the proper taking of life and property protection measures. The present work intends to deepen the analysis of meteorological data with ML techniques to improve our capacity in short-range forecasting of rainfall. To this end, relationships between thermodynamic parameters derived by satellite measurements and recorded rainfall by in situ gauges, along with outputs from a numerical atmospheric model are analyzed. The main purpose of the work is to find the best relationships between the atmospheric conditions and the formation of clouds that lead to production of rainfall and build a ML model for nowcasting of rainfall. Several ML methods are used, i.e., Auto Regression, Ensemble Machine Learning, and Deep Learning, and their results are compared in order to find the best fit model.

Keywords: machine learning; meteorological data; rainfall prediction; random forest; deep learning

Citation: Kyros, G.; Manolas, I.; Diamantaras, K.; Dafis, S.; Lagouvardos, K. A Machine Learning Approach for Rainfall Nowcasting Using Numerical Model and Observational Data. *Environ. Sci. Proc.* **2023**, *26*, 11. https://doi.org/10.3390/environsciproc2023026011

Academic Editors: Konstantinos Moustris and Panagiotis Nastos

Published: 23 August 2023

1. Introduction

Nowadays, developing technology and the rise in the volume of information make data analysis a complex process. The evolving challenges of handling the volume and complexity of Big Data have led to several studies focusing on machine-learning models [1]. Weather forecasting is based on large datasets derived from radars, satellites, and ground-based measurements, thus requiring efficient data storage, processing, and data mining technologies [2].

Rainfall affects many aspects of human activities, such as agricultural production, construction, power generation, forestry, and tourism, among others [3]. Between 1970 and 2019, 44% of all disasters and 31% of all economic losses worldwide were related to extreme rainfall and floods [4]. Floods and similar high-impact weather events are expected to occur more frequently worldwide in the coming years. According to the latest IPCC report, the intensity of extreme rainfall events has increased in recent years in many regions, including most areas of Europe [5]. Particularly in Greece, rainfall is a very significant phenomenon, as its intensity in the past has caused unprecedented disasters and loss of life during catastrophic floods [6–8]. Accordingly, the accurate prediction of rainfall remains a crucial challenge due to its consequences. Several atmospheric factors affect the occurrence and intensity of rainfall. Temperature, humidity, solar radiation, atmospheric pressure, and cloud microphysics are some of the factors that affect the occurrence of rainfall and its intensity [9].

In this work, we applied machine learning models, using reanalysis model data, satellite data, and ground observations from meteorological stations, to forecast rainfall occurrence over a period of 3 h.

2. Materials and Methods

2.1. Study Area and Data

The study area is Western Macedonia, located in Northern Greece (Figure 1a). It is a mountainous area of 9.451. Reanalysis data were retrieved from the Copernicus Climate Database (ERA5) [10], and satellite data from GRIDSAT-B1 of NOAA [11]. Finally, the data of 5 weather stations were used from the NOAAN Network of Automatic Meteorological Stations of the National Observatory of Athens/METEO [12] (Figure 1b).

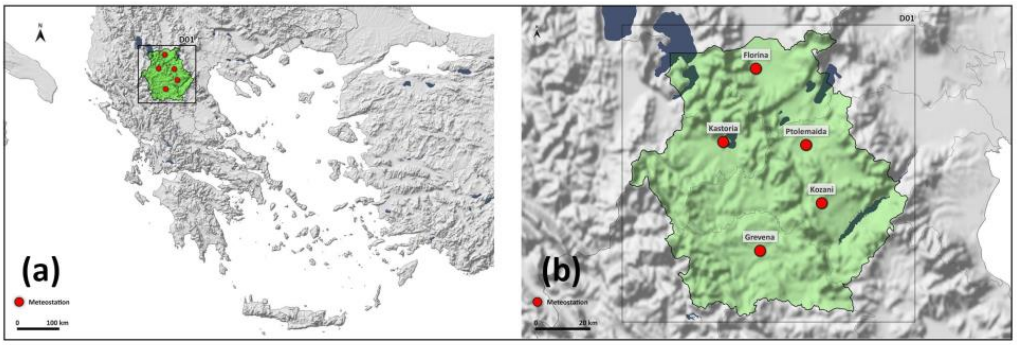

Figure 1. Location of the study area in Greece (**a**) and zoomed domain with the five selected weather stations (**b**).

The selection of 5 of 38 stations which are installed in Western Macedonia [12] was based on data availability in years from 2017 to 2020, the geographical distribution of the stations in the regional units, and the adjacency with population density. These five weather stations are located close to the five biggest cities of Western Macedonia. More specifically, the four stations (Grevena, Kastoria, Kozani, Florina) are the capitals of each regional unit, respectively.

2.1.1. Dataset Description

The dataset consists of the ERA5 reanalysis dataset; data on pressure (20 levels) and single levels were retrieved, including data on temperature, relative humidity, potential vorticity, and wind. From the GRIDSAT-B1 dataset, brightness temperature at the window channel 10.8 μm and water vapor channel 6.2 μm were utilized. Data from the NOAAN weather stations included temperature, relative humidity, atmospheric pressure, wind speed and direction, and precipitation rate. Data sources and their characteristics are shown on Table 1.

Table 1. Data sources and their characteristics.

Source	Type	Geographical Area	Date Range	Sample Rate
ERA5 (Pressure and Single levels)	Gridded 0.25°	19° E, 23° E, 23° N, 23° N	2017–2020	3 h
Satellite Data (GRIDSAT-B1)	Gridded 0.07°	19° E, 23° E, 23° N, 23° N	2017–2020	3 h
Weather Station Data (N.O.A/meteo.gr)	Dataframe	Grevena, Kastoria, Kozani, Ptolemaida, Florina	2017–2020	10 min/resampled to 3 h

2.1.2. Data Preprocessing

In order to carry out the study, a set of operations was performed to prepare the data similar to [13]. First of all, we had to extract point data on the nearest grid point to

the location of weather stations from gridded data (ERA5 and GRIDSAT-B1). All three data sources had different sample rates; thus, the next step was to resample the data to a common timestamp, namely a 3 h rate. The data from weather stations were resampled from 10 min intervals to 3 h intervals with aggregation methods (sum, mean), which were applied 10 min before and after the 3 h timestep.

Missing Values and Detection of Outliers

We had a small number of missing values, approximately 10–15 timesteps, which is almost 0.001% of all data. Thus, the missing values affected the method of interpolation [14]. For the detection of outliers, the "Z-score" method [15] was used, and all samples with a score of Z > 3 were discarded. As a result, only 0.002% of the data were removed; thus, we have a final number of 10.960 samples.

Data Normalization

Data normalization was applied only for the data of the Neural Network of LSTM. We used the min–max type (Equation (1)) [16]. Thus, all the variable values were normalized between 0 and 1. In this way, features taking values of significant volumes having the highest impact on the application of machine learning algorithms were avoided.

$$z = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

where $min(x)$ and $max(x)$ are the minimum and maximum value, respectively. x is the value to be scaled, and z is the scaled value. After preprocessing, we applied inverse transformation of normalized values back to actual values for the prediction and the error computation.

2.2. Machine Learning Models and Methods

The machine learning models which were applied to this study are Random Forest, XGBoost, and Neural Networks (LSTM). Random forest (RF) is a supervised machine learning ensemble algorithm which is based on decision tree predictors [17]. It is applied for both classification and regression applications. XGBoost, is a scalable machine learning system for tree boosting. It is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model [18,19]. Long short-term memory (LSTM) is an artificial neural network used in the fields of artificial intelligence and deep learning. An LSTM is a variant of the neural network-based modeling approach, an upgraded version of RNNs capable of learning long-term dependence that exists at various steps in the sequential time series data [20].

Each model employs all the data from three sources (numerical, satellite, and in situ data) divided into a training set, a validation set, and a test set. The target of training models was to predict the instantaneous rainfall for the next 3 h at each station.

2.2.1. Training and Test Data

To deploy, and evaluate the model, the dataset was split into three parts: 70% to the training set, 10% to the validation set, and 20% to the test set. We chose to use the split method of 70–10–20 because having too many or too few samples in the training set had a negative effect on the estimated model performance, suggesting that it is better to have a good balance between the sizes of the training set and validation set to have a reliable estimation of model performance [21]. Therefore, the training set is the portion of data used to train the model. The model should observe and learn from the training set, optimizing any of its parameters. The validation set is used to avoid overfitting. Then, once the parameters of the models have been defined, the test set was used for the final model evaluation [22]. Finally, the range of the training and validation set was from April 2017 to January 2020, and the remaining samples (February 2020 to December 2020) were used for the test set.

2.2.2. Evaluation Metrics

This section defines the metrics used to be able to evaluate the results of the algorithms used. The performance of each model was evaluated and compared using two different metrics: Mean Absolute Error (MAE) (Equation (2)), and Root-Mean-Square Error (RMSE) (Equation (3)).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

where i is the number of samples, y_i the observed value, $\hat{y}_i$ the predicted value, and n is the total number of samples [23].

3. Results

The evaluation metrics were calculated for both Gradient Boosting, Random Forest, and Neural Network model, after they were trained. Results are shown on Table 2.

Table 2. Metric scores of predicted instantaneous rainfall (>0.2 mm) for the next 3 h for the 3 trained models, evaluated using the testing dataset for 5 stations.

Model	MAE [mm]	RMSE [mm]
Random Forest	0.4	0.6
XGBoost	0.5	0.7
LSTM	0.7	1.1

The Machine Learning model results were validated by performing a cross-validation and rebuilding the models. We can see that Random Forest outperformed the other two models. Second in ranking was the XGBoost regressor and last was the Neural Networks (LSTM). This ranking can be explained from data exploration. Our problem seems to fit better on regression algorithms than on Neural Networks; thus, the Random Forest regressor and XGBoost regressor predicted better results [19,24]. Moreover, the Random Forest predicted better results than other models on high values of rainfall. The Recurrent Neural Networks and LSTM had poor predictive accuracy on high values but seem to perform better on lower values. Finally, the XGBoost had a similar performance to Random Forest but was worse in high and extreme values of rainfall. The best overall results were from the station of Grevena using Random Forest as shown in Figure 2a,b, with the best result of RMSE of 0.49.

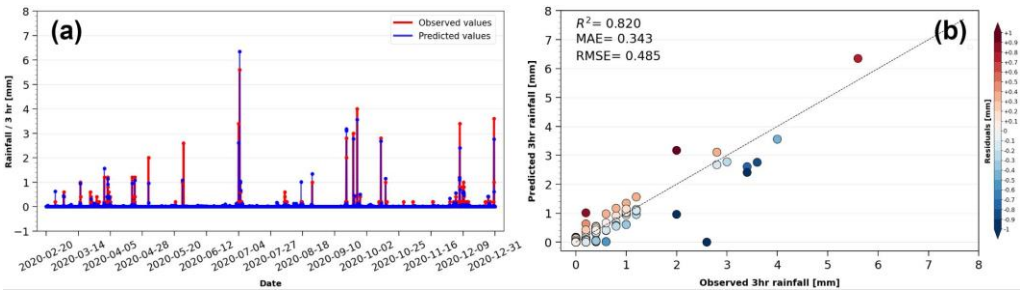

Figure 2. Comparison of actual and predicted 3 h instantaneous rainfall plots (**a**) and (**b**) in the station of Grevena with results from the Random Forest model.

4. Discussion and Conclusions

This study developed machine learning models for rainfall nowcasting for the next 3 h. The analysis was based on 4 years of data from meteorological stations over the study area, located in Western Macedonia, Greece. In general, all the models produced consistent

predictions and learned the complex relationship between atmospheric conditions and production of rainfall. The results motivate further research into the application of machine learning models for forecasting rainfall and complementing the existing numerical weather prediction methods. Finally, we plan to use baseline models to compare them with our proposed models.

Author Contributions: Conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, G.K. and I.M.; Conceptualization, project administration, resources, supervision, writing—review and editing, S.D.; Project administration, resources, supervision, K.L.; Supervision, K.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of some data. Data were obtained from the Institute of Environmental Research and Sustainable Development, National Observatory of Athens, National Centers for Environmental Information of NOAA and Copernicus Climate Database of ECMWF.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Olawoyin, A.M.; Leung, C.K.; Hryhoruk, C.C.J.; Cuzzocrea, A. Big Data Management for Machine Learning from Big Data. In Proceedings of the 37th International Conference on Advanced Information Networking and Applications (AINA-2023), Juiz de Fora, Brazil, 29–31 March 2023; pp. 393–405. [CrossRef]
2. Hussein, E.; Sadiki, R.; Jafta, Y.; Sungay, M.M.; Ajayi, O.; Bagula, A. Big Data Processing Using Hadoop and Spark: The Case of Meteorology Data. In Proceedings of the 11th EAI International Conference, AFRICOMM 2019, Porto-Novo, Benin, 3–4 December 2019; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. 2020; pp. 180–185. [CrossRef]
3. World Health Organization: Climate Change and Human Health: Risks and Responses. World Health Organization, January 2003. Available online: https://apps.who.int/iris/handle/10665/42749 (accessed on 25 April 2023).
4. WMO Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2019) (WMO-No. 1267). Available online: https://library.wmo.int/index.php?lvl=notice_display&id=21930#.ZGJGDs5ByUm (accessed on 26 April 2023).
5. IPCC. *Climate Change 2022: Impacts, Adaptation and Vulnerability*; Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Pörtner, H.-O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., et al., Eds.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2022; 3056p. [CrossRef]
6. Papagiannaki, K.; Diakakis, M.; Kotroni, V.; Lagouvardos, K.; Andreadakis, E. Hydrogeological and Climatological Risks Perception in a Multi-Hazard Environment: The Case of Greece. *Water* **2019**, *11*, 1770. [CrossRef]
7. Papagiannaki, K.; Kotroni, V.; Lagouvardos, K.; Ruin, I.; Bezes, A. Urban Area Response to Flash Flood–Triggering Rainfall, Featuring Human Behavioral Factors: The Case of 22 October 2015 in Attica, Greece. *Weather. Clim. Soc.* **2017**, *9*, 621–638. [CrossRef]
8. Kotroni, V.; Lagouvardos, K.; Bezes, A.; Dafis, S.; Galanaki, E.; Giannaros, C.; Giannaros, T.; Karagiannidis, A.; Koletsis, I.; Kopania, T.; et al. Storm Naming in the Eastern Mediterranean: Procedures, Events Review and Impact on the Citizens Risk Perception and Readiness. *Atmosphere* **2021**, *12*, 1537. [CrossRef]
9. Liyew, C.M.; Melese, H.A. Machine Learning Techniques to Predict Daily Rainfall Amount. *Res. Sq.* **2021**. [CrossRef]
10. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 Global Reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
11. Knapp, K.R.; Ansari, S.; Bain, C.L.; Bourassa, M.A.; Dickinson, M.J.; Funk, C.; Helms, C.N.; Hennon, C.C.; Holmes, C.D.; Huffman, G.J.; et al. Globally Gridded Satellite Observations for Climate Studies. *Bull. Am. Meteorol. Soc.* **2011**, *92*, 893–907. [CrossRef]
12. Lagouvardos, K.; Kotroni, V.; Bezes, A.; Koletsis, I.; Kopania, T.; Lykoudis, S.; Mazarakis, N.; Papagiannaki, K.; Vougioukas, S. The Automatic Weather Stations NOANN Network of the National Observatory of Athens: Operation and Database. *Geosci. Data J.* **2017**, *4*, 4–16. [CrossRef]
13. Iliou, T.; Anagnostopoulos, C.-N.; Nerantzaki, M.; Anastassopoulos, G. A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS), Rhodes Island, Greece, 25–28 September 2015. [CrossRef]

14. Somasundaram, R.S.; Nedunchezhian, R. Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values. *Int. J. Comput. Appl.* **2011**, *21*, 14–19. [CrossRef]
15. Taylor, J.K.; Cihon, C. *Statistical Techniques for Data Analysis*; Taylor & Francis Group: New York, NY, USA, 2004. [CrossRef]
16. Patro, S.G.K.; Sahu, K.K. Normalization: A Preprocessing Stage. *arXiv* **2015**, arXiv:1503.06462. [CrossRef]
17. Meenal, R.; Michael, P.A.; Pamela, D.; Rajasekaran, E. Weather Prediction Using Random Forest Machine Learning Model. *Indones. J. Electr. Eng. Comput. Sci.* **2021**, *22*, 1208. [CrossRef]
18. Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]
19. Ibrahem Ahmed Osman, A.; Najah Ahmed, A.; Chow, M.F.; Feng Huang, Y.; El-Shafie, A. Extreme Gradient Boosting (Xgboost) Model to Predict the Groundwater Levels in Selangor Malaysia. *Ain Shams Eng. J.* **2021**, *12*, 1545–1556. [CrossRef]
20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
21. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef] [PubMed]
22. Omary, Z.; Mtenzi, F. Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning. *Int. J. Infonomics* **2010**, *3*, 314–325. [CrossRef]
23. Liu, Y.; Zhou, Y.; Wen, S.; Tang, C. A Strategy on Selecting Performance Metrics for Classifier Evaluation. *Int. J. Mob. Comput. Multimed. Commun.* **2014**, *6*, 20–35. [CrossRef]
24. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M.C. Modeling Spatial Patterns of Fire Occurrence in Mediterranean Europe Using Multiple Regression and Random Forest. *For. Ecol. Manag.* **2012**, *275*, 117–129. [CrossRef]