UNIVERSIDAD
Ð CÓRDOBA

TESIS DOCTORAL POR COMPENDIO DE PUBLICACIONES

PROGRAMA DE DOCTORADO:

“RECURSOS NATURALES Y GESTIÓN SOSTENIBLE”

MODELOS INTELIGENTES PARA LA MEJORA DE ESTIMACIONES Y PREDICCIONES AGROMETEOROLÓGICAS

SMART MODELS TO IMPROVE AGROMETEOROLOGICAL ESTIMATIONS AND PREDICTIONS

Autor:
Juan Antonio Bellido Jiménez
Directores:
Dr. Javier Estévez Gualda
Dra. Amanda Penélope García Marín

TITULO: *Smart models to improve agrometeorological estimations and predictions*

AUTOR: *Juan Antonio Bellido Jiménez*

© Edita: UCOPress. 2023
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

TÍTULO DE LA TESIS: Modelos inteligentes para la mejora de estimaciones y predicciones agrometeorológicas

DOCTORANDO/A: Juan Antonio Bellido Jiménez

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

Desde que en enero de 2020 el doctorando Juan Antonio Bellido Jiménez iniciara su Tesis doctoral, su evolución en términos científicos, formativos y en otras competencias del ámbito investigador, ha sido excelente. Desde 31 de enero de 2020 se integró en el Equipo de Trabajo del Proyecto Nacional de Investigación Retos AGL2017-87658-R Predicción inteligente de la variabilidad espacio-temporal de la aridez en el Sur de España, contribuyendo a conseguir los objetivos planteados en las distintas fases.

Su grado de autonomía y motivación científica son puntos clave a resaltar dentro de este periodo predoctoral, demostrando una alta capacidad para abordar las distintas tareas de su investigación. Desde el análisis conceptual de las hipótesis hasta la programación de los modelos inteligentes, el doctorando ha profundizado en técnicas innovadoras para aplicarlas al ámbito de la Meteorología Agrícola, obteniendo excelentes resultados y generando repositorios de acceso libre de gran impacto para la comunidad científico-técnica. Además, el doctorando ha desarrollado una gran capacidad para manejar y gestionar grandes volúmenes de datos, así como para optimizar los procesos de entrenamiento y validación de los modelos que se han generado en esta Tesis doctoral. También ha aprendido a realizar una amplia labor de consulta bibliográfica continua como requisito previo a cualquier problema científico que se pretenda abordar y ha realizado todas las actividades de su plan formativo de forma excelente. Los objetivos del plan de investigación se han conseguido ampliamente, conformando una tesis doctoral de altísima calidad en el ámbito internacional ya que contribuye de forma notable a la mejora de las predicciones y estimaciones que los modelos agrometeorológicos actuales realizan.

Finalmente, se han derivado 6 trabajos publicados en revistas de impacto dentro del Journal Citation Report (JCR/SCI), estando 4 de ellas en el primer cuartil (Q1), una en el segundo (Q2) y otra en el tercero (Q3):

- Estévez, J; Bellido-Jiménez, JA; Liu, X; García-Marín, AP (2020). Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment. Water (12). https://doi.org/10.3390/w12071909 . JCR: 3.103 (2020). 36/100 (Q2) – Water Resources.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2021). New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agricultural Water

Management, (245). https://doi.org/10.1016/j.agwat.2020.106558 . JCR: 6.611 (2021). 5/90 (Q1) – Agronomy.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2021). Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions. Applied Energy, (298). https://doi.org/10.1016/j.apenergy.2021.117211 . JCR: 11.446 (2021). 15/119 (Q1) – Energy & Fuels.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2021). Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. Atmosphere, (12). https://doi.org/10.3390/atmos12091158 . JCR: 3.110 (2021). 59/94 (Q3) – Meteorology & atmosphere sciences.

- Bellido-Jiménez, JA; Estévez, J; Vanschoren, J; García-Marín, AP (2022). AgroML: An Open-Source Repository to Forecast Reference Evapotranspiration in Different Geo-Climatic Conditions Using Machine Learning and Transformer-Based Models, (12). https://doi.org/10.3390/agronomy12030656 . JCR: 3.949 (2021). 18/90 (Q1) – Agronomy.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2022). Regional machine learning method to outperform temperature-based reference evapotranspiration estimations in Southern Spain. Agricultural Water Management, (274). https://doi.org/10.1016/j.agwat.2022.107955 . JCR: 6.611 (2021). 5/90 (Q1) - Agronomy

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 23 de marzo de 2023

Firma del/de los director/es

Javier Estévez

Fdo.: Javier Estévez Gualda

Fdo.: Amanda P. García Marín

TESIS POR COMPENDIO DE ARTÍCULOS

La presente tesis doctoral cumple con los requisitos establecidos en la normativa del programa de doctorado en Recursos Naturales y Gestión Sostenible de la Universidad de Córdoba para su presentación como compendio de publicaciones. Este compendio está formado por 6 artículos publicados en revistas incluidas en los tres primeros cuartiles de la última relación publicada por el Journal Citation Report (JCR). Además, el doctorando aparece como primer autor de las publicaciones, habiendo aceptado los coautores su presentación en esta tesis doctoral

- Estévez, J; Bellido-Jiménez, JA; Liu, X; García-Marín, AP (2020). Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment. Water (12). https://doi.org/10.3390/w12071909. **JCR: 3.103** (2020). 36/100 (Q2) – Water Resources.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2021). New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agricultural Water Management, (245). https://doi.org/10.1016/j.agwat.2020.106558. **JCR: 6.611** (2021). 5/90 (Q1) – Agronomy.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2021). Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions. Applied Energy, (298). https://doi.org/10.1016/j.apenergy.2021.117211. **JCR: 11.446 (2021)**. 15/119 (Q1) – Energy & Fuels.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2021). Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. Atmosphere, (12). https://doi.org/10.3390/atmos12091158. **JCR:3.110** (2021). 59/94 (Q3) – Meteorology & atmosphere sciences.

- Bellido-Jiménez, JA; Estévez, J; Vanschoren, J; García-Marín, AP (2022). AgroML: An Open-Source Repository to Forecast Reference Evapotranspiration in Different Geo-Climatic Conditions Using Machine Learning and Transformer-Based Models, (12). https://doi.org/10.3390/agronomy12030656. **JCR: 3.949** (2021). 18/90 (Q1) – Agronomy.

- Bellido-Jiménez, JA; Estévez, J; García-Marín, AP (2022). Regional machine learning method to outperform temperature-based reference evapotranspiration estimations in Southern Spain. Agricultural Water Management, (274). https://doi.org/10.1016/j.agwat.2022.107955. **JCR: 6 611** (2021). 5/90 (Q1) - Agronomy

Córdoba, marzo 2023

El doctorando

Fdo: Juan Antonio Bellido Jiménez

Agradecimientos

En una ocasión vi cómo comparaban la realización de una tesis doctoral con la subida al pico de una montaña. No podría ser más acertada, sin embargo, habían excluido el fundamental trabajo del guía de la expedición, el director de la tesis; quien te anima, enseña y orienta hacia el mejor camino. El ascenso es inevitable, pero el guía experimentado consigue focalizar y distribuir los esfuerzos evitando el agotamiento que producen las paredes verticales. Es por lo que agradezco a mis directores, Amanda y Javier, el apoyo, orientación y confianzas depositada y trasladada durante la realización de toda la tesis. Gracias.

Así mismo, de forma particular, agradezco a Javier sus consejos, no solo durante los años de la tesis, sino en los previos y, sobre todo, su apoyo incondicional. Gracias a ti, hoy, estoy terminando esta tesis que espero cumpla tus expectativas. También deseo que sea el punto de partida de años venideros cargados de éxitos, tanto profesionales como musicales.

Agradezco a Joaquín Vanschoren el haberme brindado la oportunidad de compartir estancia con su grupo de investigación en Eindhoven; tres maravillosos meses en los que, además de aprender, tuve el honor de conocer a increíbles personas, entre las que se encuentran Israel, Ceren, Onur, Bilge, Prabhant, Pieter, Wouter, Sibylle, Andrei y Loek.

También agradezco la constancia, dedicación, saber hacer y bonhomía de Ezequiel, antiguo compañero de trabajo, así como a Carlos, que depositó en mi una confianza que me hizo crecer y realizarme como profesional. Cómo olvidar con quienes he crecido y seguiré haciéndolo, Germán, compañero de universidad y trabajo, Gregorio, amigo de toda la vida, quien además me ha ayudado con esta tesis, revisando códigos y asesorando.

Tampoco puedo olvidar a mi luchadora favorita; madre, Sergito y yo siempre estaremos, no lo olvides. A mi padre, confidente, apoyo y ejemplo a seguir; te quiero. A mi hermano Sergio, al que admiro por su forma de ser, su tesón y perseverancia; sé que disfrutaremos y celebraremos muchos de tus logros. A Victoria, Jesús y Vito, quienes me habéis enseñado nuevas acepciones de la palabra familia y, como ya dijo mi hermano, el mayor descubrimiento de los últimos años. Me siento muy afortunado y agradecido de teneros a mi lado. Os quiero.

A Almu, mi compañera de vida, quien más ha sufrido este sacrificio por aventurarme a realizar la tesis doctoral y dedicarme a la investigación. Te quiero.

Resumen

La población mundial, en continuo crecimiento, alcanzará de forma estimada los 9,7 mil millones de habitantes en el 2050. Este incremento, combinado con el aumento en los estándares de vida y la situación de emergencia climática (aumento de la temperatura, intensificación del ciclo del agua, etc.) nos enfrentan al enorme desafío de gestionar de forma sostenible los cada vez más escasos recursos disponibles. El sector agrícola tiene que afrontar retos tan importantes como la mejora en la gestión de los recursos naturales, la reducción de la degradación medioambiental o la seguridad alimentaria y nutricional. Todo ello condicionado por la escasez de agua y las condiciones de aridez: factores limitantes en la producción de cultivos. Para garantizar una producción agrícola sostenible bajo estas condiciones, es necesario que todas las decisiones que se tomen estén basadas en el conocimiento, la innovación y la digitalización de la agricultura de forma que se garantice la resiliencia de los agroecosistemas, especialmente en entornos áridos, semi-áridos y secos sub-húmedos en los que el déficit de agua es estructural.

Por todo esto, el presente trabajo se centra en la mejora de la precisión de los actuales modelos agrometeorológicos, aplicando técnicas de inteligencia artificial. Estos modelos pueden proporcionar estimaciones y predicciones precisas de variables clave como la precipitación, la radiación solar y la evapotranspiración de referencia. A partir de ellas, es posible favorecer estrategias agrícolas más sostenibles, gracias a la posibilidad de reducir el consumo de agua y energía, por ejemplo. Además, se han reducido el número de mediciones requeridas como parámetros de entrada para estos modelos, haciéndolos más accesibles y aplicables en áreas rurales y países en desarrollo que no pueden permitirse el alto costo de la instalación, calibración y mantenimiento de estaciones meteorológicas automáticas completas. Este enfoque puede ayudar a proporcionar información valiosa a los técnicos, agricultores, gestores y responsables políticos de la planificación hídrica y agraria en zonas clave.

Esta tesis doctoral ha desarrollado y validado nuevas metodologías basadas en inteligencia artificial que han ser vido para mejorar la precision de variables cruciales en al ámbito agrometeorológico: precipitación, radiación solar y evapotranspiración de referencia. En particular, se han modelado sistemas de predicción y rellenado de huecos de precipitación a diferentes escalas utilizando redes neuronales. También se han desarrollado modelos de estimación de radiación solar utilizando exclusivamente parámetros térmicos y validados en zonas con características climáticas similares a lugar de entrenamiento, sin necesidad de estar geográficamente en la misma región o país. Analógamente, se han desarrollado modelos de estimación y predicción de evapotranspiración de referencia a nivel local y regional utilizando también solamente datos de temperatura para todo el proceso: regionalización, entrenamiento y validación. Y finalmente, se ha creado una librería de Python de código abierto a nivel internacional (AgroML) que facilita el proceso de desarrollo y aplicación de modelos de inteligencia

artificial, no solo enfocadas al sector agrometeorológico, sino también a cualquier modelo supervisado que mejore la toma de decisiones en otras áreas de interés.

Abstract

The world population, which is constantly growing, is estimated to reach 9.7 billion people in 2050. This increase, combined with the rise in living standards and the climate emergency situation (increase in temperature, intensification of the water cycle, etc.), presents us with the enormous challenge of managing increasingly scarce resources in a sustainable way. The agricultural sector must face important challenges such as improving natural resource management, reducing environmental degradation, and ensuring food and nutritional security. All of this is conditioned by water scarcity and aridity, limiting factors in crop production. To guarantee sustainable agricultural production under these conditions, it is necessary to based all the decision made on knowledge, innovation, and the digitization of agriculture to ensure the resilience of agroecosystems, especially in arid, semi-arid, and sub-humid dry environments where water deficit is structural.

Therefore, this work focuses on improving the precision of current agrometeorological models by applying artificial intelligence techniques. These models can provide accurate estimates and predictions of key variables such as precipitation, solar radiation, and reference evapotranspiration. This way, it is possible to promote more sustainable agricultural strategies by reducing water and energy consumption, for example. In addition, the number of measurements required as input parameters for these models has been reduced, making them more accessible and applicable in rural areas and developing countries that cannot afford the high cost of installing, calibrating, and maintaining complete automatic weather stations. This approach can help provide valuable information to technicians, farmers, managers, and policy makers in key water and agricultural planning areas.

This doctoral thesis has developed and validated new methodologies based on artificial intelligence that have been used to improve the precision of crucial variables in the agrometeorological field: precipitation, solar radiation, and reference evapotranspiration. Specifically, prediction systems and gap-filling models for precipitation at different scales have been modeled using neural networks. Models for estimating solar radiation using only thermal parameters have also been developed and validated in areas with similar climatic characteristics to the training location, without the need to be geographically in the same region or country. Similarly, models for estimating and predicting reference evapotranspiration at the local and regional level have been developed using only temperature data for the entire process: regionalization, training, and validation. Finally, an internationally open-source Python library (AgroML) has been created to facilitate the development and application of artificial intelligence models, not only focused on the agrometeorological sector but also on any supervised model that improves decision-making in other areas of interest.

Index

Agradecimientos 8

Resumen 10

Abstract 12

Index 14

List of Tables 18

List of Figures 24

List of acronyms 30

Lyst of symbols 32

Chapter 1. Introduction 34

- 1. The current situation 34
- 2. Agrometeorological variables 35
  - 2.1. Air Temperature 35
  - 2.2. Relative humidity 36
  - 2.3. Solar radiation 37
  - 2.4. Wind speed 39
  - 2.5. Precipitation 39
  - 2.6. Reference evapotranspiration 39
- 3. Automated Weather Stations in Andalusia 41
- 4. Quality Control Procedures 42
- 5. Feature engineering and selection 43
- 6. Machine learning models 45
  - 6.1. Linear Regression 45
  - 6.2. Multilayer Perceptron (MLP) 45
  - 6.4. Support Vector Machine (SVM) 47
  - 6.5. Random Forest (RF) 49
- 7. Hyperparameter selection 50
- 8. AgroML 51
- 9. Hypothesis 51
- 10. Objectives 52
- 11. References 52

Chapter 2. Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment 62

- 1. Introduction 63
- 1.1. Wavelet Multiscale Analysis 64

1.2. Availability of Short-Term Meteorological Series 65
2. Materials and methods 66
2.1. Source of data 66
2.2. Development of Wavelet Neural Network (WNN) Models 70
2.3. Statistical Analysis and Performance Criteria 75
3. Results and Discussion 76
3.1. Pre-Processing Input Datasets 76
3.2. Performance of the Models 76
4. Conclusions 83
5. References 84
Chapter 3. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain 92
1. Introduction 93
2. Materials and methods 95
2.1. Source of data 95
2.2. Quality assurance procedures 99
2.3. FAO56 – PM equation 99
2.4. Temperature-based methods 100
2.5. Machine learning models 102
2.6. Data standardization 111
2.7. Statistical analysis 111
3. Results and Discussion 111
3.1. Performance of models in the different locations 112
3.2. Performance of the configurations 116
3.3. Seasonal performance behavior 120
3.4. Overall discussion 121
4. Conclusions 122
5. References 123
Chapter 4. Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions 132
1. Introduction 133
1.1. About the related work 133
1.2. Research gaps and scientific contribution 135
2. Theoretical overview 136
2.1. Empirical temperature-based models 137
2.2. Machine Learning models 137
2.2.5. Random Forest (RF) 140

3. Materials and methods 140
3.1. Source of data 140
3.2. Quality assurance procedures 146
3.3. Bayesian optimization 146
3.4. Data standardization 147
3.5. Statistical analysis 147
4. Results and Discussion 148
4.1. Models performance in different locations 148
4.2. Performance of the configurations 154
4.3. Seasonal performance 158
4.4. Models' application to new sites 160
4.5. Overall discussion 161
5. Conclusions and future work 162
6. References 163
Chapter 5. Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain 172
1. Introduction 172
2. Materials and Methods 175
2.2. Methodology 177
2.3. Multilayer Perceptron (MLP) 180
2.4. Support Vector Machine (SVM) 181
2.5. Random Forest (RF) 181
2.6. Bayesian Optimization 182
2.7. Evaluation Metrics 185
3. Results and Discussion 186
3.1. Using Neighbor Stations 186
3.2. Using Data from the Target Station 189
3.3. Comparison of the Two Areas 191
3.4. Seasonality Performance 193
3.5. General Discussion 196
4. Conclusions 197
References 197
Chapter 6. AgroML: an open-source repository to forecast reference evapotranspiration in different geo-climatic conditions using machine learning and Transformer-based models 203
1. Introduction 204
2. Materials and Methods 206
2.1. Study area and dataset 206

2.2. Preprocessing methodology 211
2.3. Reference evapotranspiration calculation 213
2.4. Baselines 214
2.5. Machine learning models 214
2.6. Bayesian optimization 219
2.7. Evaluation metrics 226
3. Results and Discussion 226
3.1. Baseline performance 227
3.2. Analysis of ML performance 228
3.3. Assessing the different configurations 238
3.4. Overall discussion 240
4. Conclusions 244
5. References 245
Chapter 7. A regional machine learning method to outperform temperature-based reference evapotranspiration estimations in Southern Spain 250
1. Introduction 251
2. Materials and methods 254
2.1. Source of data and data management 254
2.2. Reference evapotranspiration 263
2.3. Machine learning models 264
2.3.2. Extreme Learning Machine 265
2.4. Bayesian optimization 265
2.5. Statistical analysis 267
3. Results and discussion 268
3.1. The evaluation of ML models in the different regions 268
3.2. Assessment of configurations 273
3.4 Overall discussion 279
4. Conclusions 281
5. References 281
Chapter 8. Conclusions 288
Chapter 9. Future works 290

List of Tables

Table 1. 1. Sensors specifications from the RIA AWS network [31]........................... 42

Table 2. 1. Name of the station, province, coordinates, elevation and data time-period of the weather stations used in this study (Southern Spain)……………………………..….68

Table 2. 2. Statistics of monthly precipitation, maximum and minimum temperature (Std: Standard Deviation; Max: Maximum; Min: Minimum)... 71

Table 2. 3. Inputs and the number of variables of each of the wavelet neural network models (WNN) models evaluated in this work (i = month; MOY = month of the year; P = precipitation; DTR_M = mean diurnal temperature range; DTR_X = maximum diurnal temperature range; DTR_N = minimum diurnal temperature range; MTR=monthly temperature range; T_X=maximum temperature; Tn = minimum temperature). 74

Table 2. 4. Summary of correlation coefficient (R), root mean square error (RMSE), mean absolute percentage error (MAPE) and Nash–Sutcliffe model efficiency coefficient (NSE) values for all the models assessed. 79

Table 3. 1. Summary of weather stations sites used in this paper (Lat: Latitude; Lon: Longitude; Altitude: elevation above mean sea level, mean annual precipitation, and Aridity Index as the relation Precipitation/ET_0 as the annual mean of the whole dataset period)……………………………………………………………………………………..96

Table 3. 2. Statistics of maximum and minimum temperature, and ET_0 (Max: Maximum, Min: Minimum, Std: Standard Deviation. 98

Table 3. 3. RIA automated weather stations sensor specifications............................. 100

Table 3. 4. Correction of dew temperature (T_{dew}) estimation from minimum temperature (Tn) proposed by Todorovic et al. (2013). 101

Table 3. 5. Input climatic parameter configurations (T_x: Maximum temperature, T_n: Minimum temperature, T_{dew}: dew point temperature, R_a: Extraterrestrial radiation, e_s: mean saturation vapor pressure, e_a: actual vapor pressure, VPD: vapor pressure deficit, $Energy_T$: the integral of the daily thermal signal, Hourmin is the difference between the time when sunset occurs and the time the maximum temperature occurs). 111

Table 3. 6. Minimum (Min), mean, and maximum (Max) values of MBE, RMSE, R^2, and NSE for all the stations using different models: Hargreaves-Samani (HS), Calibrated Hargreaves-Samani (HSMOD), Multilayer Perceptron (MLP), Extreme Learning Machines (ELM), Support Vector Machines (SVM), Generalised Regression Neural Network (GRNN), XGBoost and Random Forest (RF). The symbol * means '*no-value*'. 113

Table 3. 7. Best NSE and RMSE (mm/day) values for all the stations (TAB=Tabernas; CON=Conil; COR=Córdoba; ARO=Aroche; MAG=Málaga) in the different input configurations (Table 4). The model in parenthesis is the model that obtained the best NSE and RMSE, respectively... 119

Table 3. 8. Seasonal mean NSE and RMSE values of all the models and stations in the different input configurations (table 4).. 121

Table 4. 1 Summary of automated weather station sites used in this paper (Lat: Latitude; Lon: Longitude; Altitude: elevation above mean sea level and Aridity Index as the relation Precipitation/ET_0 as the annual mean of the whole dataset period)……………………143

Table 4. 2. Specifications of the RIAA sensors 145

Table 4. 3. Input configurations of all the models, where T_X is the maximum daily temperature, T_n is the minimum daily temperature, R_a is the extraterrestrial solar radiation, $Energy_T$ is the integral of the daily thermal signal, $Hourmin_{Tx}$ is the difference between the time when sunset and T_X occurs, and $Hourmin_{Tn}$ is the difference between the time when sunrise and T_n occurs.. 148

Table 4. 4. Maximum (max), arithmetic mean (mean), and minimum (min) values for MBE, RMSE, NSE, and R^2, for all the locations and models (BC = Bristow and Campbell, HS = Hargreaves and Samani, ELM =Extreme Learning Machine, GRNN = Generalised Regression Neural Networks, MLP = Multilayer Perceptron, RF = Random Forest, SVM = Support Vector Machine and XGBoost = Extreme Gradient Boosting) evaluated in this work. For MBE, the minimum and maximum values are in terms of absolute values. The GPI values for each location and model were also reported, and the best mean values of all the statistics are in bold.. 151

Table 4. 5. Best NSE and RMSE (W/m^2) performances for the different configurations and stations assessed (ASH08 = Asheville, ALM04 = Tabernas. CAD07 = Jimena de la Frontera. COR06 = Córdoba. GRA03 = Loja. HUE08 = El Campillo. JAE07 = Mancha Real. MAG01 = Málaga. SEV09 = Écija) and the mean performance of all the stations. In parenthesis, it is shown the model used, while the best values are in bold............. 157

Table 4. 6. GPI values for the different models (BC – Bristow and Campbell. HS – Hargreaves Samani. MLP – Multilayer Perceptron. ELM – Extreme Learning Machine. RF – Random Forest. SVM – Support Vector Machine. XGBoost – Extreme Gradient Boost. and GRNN – Generalized Regression Neural Network) and configurations (see table 4. 3). The best values are in bold. 158

Table 4. 7. Mean NSE and RRMSE seasonal values for the different configurations assessed (table 4), as well as HS (Hargreaves Samani) and BC (Bristow Campbell). The best values are in bold. 159

Table 4. 8. Statistical results of the models from Córdoba, El Campillo, Málaga and Tabernas for new non-trained sites with similar climatic conditions (Bélmez, Cabra, Almuñécar and Huércal-Overa). as well as the non-calibrated Hargreaves-Samani results. 160

Table 5. 1. Geo-climatic characteristics of the AWS assessed in this work (lat.: latitude; long.: longitude; alt.: elevation above mean sea level)………………………………..176

Table 5. 2. . Inputs configurations of the different models and approaches assessed. DOY represents the day of year, P corresponds to precipitation, D corresponds to distance, and i represents an index to the dataset-specific day.. 178

Table 5. 3. Hyperparameter set for each model, approach, and location, after carrying out Bayesian optimi-zation, where activation represents the activation function, the optimizer represents the optimizer function, epochs represents the number of epochs, neurons. 183

Table 5. 4. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the first area (inland locations), using data from neighbor stations. The best values for each site are in bold 187

Table 5. 5. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the second area (coastal locations), using data from neighbor stations. The best values for each site are in bold. 188

Table 5. 6. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the first area (inland locations), using data from the target station in two different approaches, with the use of the previous and following day and the use of the two previous and two following days. The best values from each station are in bold. 190

Table 5. 7. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the second area (coastal locations), using data from the target station in two different approaches, with the use of the previous and following day and the use of the two previous and two following days. The best values from each station are in bold. 191

Table 5. 8. Best improvements between simple arithmetic averaging and the best ML model from each site for R^2 and RMSE. A positive value means that ML outperformed LI. 193

Table 6. 1. Geoclimatic characteristics of the locations assessed in this work (ARO – Aroche, CON – Conil de la Frontera, COR – Córdoba, MAG – Málaga and TAB – Tabernas). Time period from 2000 to 2018……………………………………………..…...207

Table 6. 2. Minimum (Min), Mean, Maximum (Max), and Standard Deviation (Std) values of all the daily parameters measured: Maximum air temperature (Tx), Mean air temperature (Tm), Minimum air temperature (Tn), Maximum relative humidity (RHx), Mean relative humidity (RHm), Minimum relative humidity (RHn), Wind speed at 2m height (u_2), Solar radiation (Rs), Reference evapotranspiration (ET_0) at each location (Aroche -ARO-, Conil de la Frontera -CON-, Córdoba -COR-, Málaga -MAG- and Tabernas -TAB-) and for the whole dataset (2000-2018). 208

Table 6. 3. Configuration table with all configurations. HT_X represents $Hourmin_{Tx}$, HTn represents $Hourmin_{Tn}$, HSs represents $Hourmin_{Sunset}$ and HSr represents $Hourmin_{Sunrise}$ 210

Table 6. 4. Hyperparameter space for all the models assessed in this work. MLP – Multilayer Perceptron, RF – Random Forest, SVR – Support Vector Regression, ELM – Extreme Learning Machine, CNN – Convolutional Neural Network, LSTM – Long Short-Term Memory, TransformerCNN – Transformer using CNN in the feed-forward layer, Transformer LSTM – Transformer using LSTM in the feed-forward layer, ... 219

Table 6. 5. Fittest hyperparameters for the best model and configuration at every location (All the hyperparameters name are the same as the variables in AgroML) 221

Table 6. 6. RMSE values for ET_0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean monthly value from the training dataset) 227

Table 6. 7. NSE values for ET_0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean daily monthly value from the training dataset) 228

Table 6. 8. MBE values for ET_0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean daily monthly value from the training dataset) 228

Table 6. 9. Minimum (Min.), mean and maximum (Max.) of NSE, RMSE, and MBE values for all locations (TAB – Tabernas, CON – Conil, COR – Córdoba, ARO – Aroche, MAG - Málaga) and models using two different lag days window (15 days and 30 days). T_CNN refers to Transformer using CNN in the feed-forward layer, while T_LSTM refers to Transformers using LSTM in this same layer 230

Table 6. 10. Mean and minimum RMSE values (mm/day) for the different configurations at each location. The format is: mean (minimum). The best values are in bold 238

Table 7. 1. Sensors' specifications (Estévez et al., 2011)…………………………………...254

Table 7. 2. Location and some characteristics of all the automated weather stations used in this work (Prov: Province; Lat: Latitude, Lon: Longitude; ALM: Almeria; CAD: Cadiz; COR: Cordoba; GRA: Granada; JAE: Jaen; HUE: Huelva; MAG: Malaga; SEV: Seville, u_2: Wind speed measured at two meters, Reg: Region). 257

Table 7. 3. All configurations and the features assessed in this work for each one. Ra (day) corresponds to Extraterrestrial Solar radiation from the same day, Tx (day-1) corresponds to maximum daily air temperature from the previous day, Tx (day) corresponds to the maximum daily air temperature from the same day, Tn (day-1) corresponds to minimum daily air temperature from the previous day, Tn (day) corresponds to minimum daily air temperature from the same day, DTR (day-1) corresponds to the difference between Tx and Tn from the previous day, DTR (day) corresponds to the difference between Tx and Tn from the same day, $Energy_T$ (day-1) represents the integral of the half-hourly temperature records from the previous day, $Energy_T$ (day-1) represents the integral of the half-hourly temperature records from the

same day, HTx (day) represents the time when Tx occurs from the same day and HTn (day) represents the time when Tn occurs from the same day 262

Table 7. 4. Hyperparameter space. MLP refers to Multilayer Perceptron, ELM refers to Extreme Learning Machine, relu refers to Rectified Linear Unit activation function, sigm to the sigmoid activation function, tanh to the hyperbolic tangent activation function, rbf_l2 a Radial Basis Function using L2 regularization 266

Table 7. 5. Statistical analysis regarding the different configurations for all the stations using MLP. The best values are in bold. RMSE represents Root Mean Square Error, NSE represents Nash-Sutcliffe Efficiency model, R^2 represents the Coefficient of Determination, MBE represents the Mean Bias Error and GPI represents the Global Performance Indicator. 274

Table 7. 6. Statistical analysis regarding the different configurations for AWSs from region 1 using MLP. RMSE represents Root Mean Square Error, NSE represents Nash-Sutcliffe Efficiency model, R^2 represents the Coefficient of Determination, MBE represents the Mean Bias Error and GPI represents the Global Performance Indicator. 276

Table 7. 7. Statistical analysis regarding the different configurations for AWSs from region 2 using MLP. RMSE represents Root Mean Square Error, NSE represents Nash-Sutcliffe Efficiency model, R^2 represents the Coefficient of Determination, MBE represents the Mean Bias Error and GPI represents the Global Performance Indicator 277

List of Figures

Figure 1. 1. The three-phase cycle of an automation system [4] 35

Figure 1. 2. Extraterrestrial solar radiation in the most northern and southern continental sites of Andalusia (El Viso and Tarifa, respectively) 37

Figure 1. 3. Solar radiation components. 38

Figure 1. 4. AWS sites from AEMET, RIA, and RAIF 41

Figure 1. 5. Semi-hourly temperature and relative humidity records on 24th May 2015 at the AWS from RIA located in Cordoba. The Tx time and Tn time represent the time at which T_X and Tn occur, respectively 44

Figure 1. 6. Example of linear regression models: a) simple linear regression and b) multiple linear regression 45

Figure 1. 7. Multilayer Perceptron architecture example for a model having 4 input features, 1 output, and 3 hidden layers with 5, 4, and 3 neurons, respectively 46

Figure 1. 8. Architecture of a single neuron of an MLP model 46

Figure 1. 9. Activation function: a) sigmoid, b) ReLU, and c) tanh 47

Figure 1. 10. Illustrative example of slack variables in SVR models 48

Figure 2. 1. Wavelet multiresolution analysis of original time series……………………65

Figure 2. 2. Geographical distribution of the automated weather stations used in this work (Andalusia region - Southern Spain). 67

Figure 2. 3. Multilayer Perceptron Neural Network architecture used in this work 73

Figure 2. 4. Original values and decomposed sub-series of monthly precipitation by wavelet transformation at Málaga station (MAG01) (2001–2019) 74

Figure 2. 5. Box-plot of the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values obtained by using the ten models (validation datasets) for all the sites studied. On each box: the red central mark=median; bottom and top edges of the box 77

Figure 2. 6. Results of the statistical performance obtained at each of the 16 locations studied: (a) R; (b) NSE; (c) RMSE; (d) MAPE. 81

Figure 2. 7. Plot of measured and forecasted monthly precipitation at four stations: Conil de la Frontera (a), Tabernas (b), Loja (c), and Sabiote (d) using Models VI, III, X and VIII, respectively. 82

Figure 3. 1. Spatial distribution of the five weather stations used in this work (Aroche, Conil de la Frontera, Córdoba, Málaga, and Tabernas)……………………………………….96

Figure 3. 2. Monthly values of temperature, relative humidity, wind speed, and solar radiation for all the weather stations (COR – Córdoba, MAG – Málaga, TAB – Tabernas, CON – Conil and ARO – Aroche). 97

Figure 3. 3. Semihourly evolution of relative humidity and temperature, as well as dew point temperature, time Tx and time sunset in Córdoba on 7th July of 2015. The colored area below the temperature is the variable called Energy_T... 102

Figure 3. 4. One neuron structure 103

Figure 3. 5. Multilayer Perceptron (MLP) architecture 103

Figure 3. 6. Activation functions. a) sigmoid, b) tanh, c) ReLU. 104

Figure 3. 7. GRNN architecture for two inputs and two outputs................................ 106

Figure 3. 8. Support vector machines for classification tasks (SVM) and regression (SVR). a) SVM, b) SVR. 107

Figure 3. 9. Kernel application types. a) Linear kernel, b) Polynomial kernel, c) Sigmoid kernel and d) RBF kernel.. 108

Figure 3. 10. An example of Bayesian optimization. 110

Figure 3. 11. Statistical performances of mean values of RMSE, MBE, R^2 and NSE for all the models evaluated using different configurations (colors) at each location (TAB=Tabernas; CON=Conil; COR=Córdoba; ARO=Aroche; MAG=Málaga). a) RMSE, b) MBE, c)R^2 and d) NSE.. 117

Figure 3. 12. Seasonal statistical performances of RMSE, MBE, R^2, and NSE for all the configurations using different models (colors) in all stations. a) RMSE, b) MBE, c) R^2 and d) NSE. 120

Figure 4. 1. Multilayer Perceptron structure………………………………………………..138

Figure 4. 2. Generalized Regression Neural Network architecture 139

Figure 4. 3. Example of SVM for a) classification and b) regression 139

Figure 4. 4. Spatial distribution of the weather stations used in this work: a) Almuñécar, Bélmez, Cabra, Córdoba, Écija, El Campillo, Huércal-Overa, Jimena de la Frontera, Loja, Málaga, Mancha Real and Tabernas (Spain) and b) Asheville (USA)........................ 142

Figure 4. 5. Monthly values of solar radiation, wind speed, relative humidity, and temperature for the weather stations assessed in this study (ASH08 = Asheville, ALM04 = Tabernas, CAD07 = Jimena de la Frontera, COR06 = Córdoba, GRA03 = Loja, HUE08 = El Campillo, JAE07 = Mancha Real, MAG01 = Málaga, SEV09 = Écija) 146

Figure 4. 6. Statistical performances for mean RRMSE (a), NSE (b), MBE (c) and R^2 (d) in all the stations assessed (ASH08 = Asheville, ALM04 = Tabernas, CAD07 = Jimena de la Frontera, COR06 = Córdoba, GRA03 = Loja, HUE08 = El Campillo, JAE07 = Mancha Real, MAG01 = Málaga, SEV09 = Écija) and using the different configuration 155

Figure 4. 7. Seasonal performance (RRMSE, NSE, MBE, and R^2) of the different models assessed (ELM – Extreme Learning Machine, RF – Random Forest, SVM – Support Vector Machine, XGB – Extreme Gradient Boost, MLP – Multilayer Perceptron, and GRNN – Generalized Regression Neural Network) in all the stations and configurations (Table 4. 3). 159

Figure 5. 1. Spatial distribution of the fourteen automated weather stations used in this work…………………………………………………………………………………176

Figure 5. 2. Methodology flowchart. 180

Figure 5. 3. One neuron control logic 181

Figure 5. 4. Support Vector Machine for classification (a) and regression (b). 181

Figure 5. 5. RMSE (a), MBE (b), and R^2 (c) values from testing dataset for all stations and models (MLP, SVM, RF, and LI), using only precipitation data from neighbor stations. 189

Figure 5. 6. RMSE, MBE, and R^2 values from testing dataset for all stations and models (MLP, SVM, and RF), using only precipitation data from the target station. A) RMSE using approach B, B) MBE using approach B, C) R^2 using approach B, D) RMSE using approach C, E) 192

Figure 5. 7. RMSE (a), MBE (b), and R^2 (c) values of the different models (SVM, MLP, and RF) in the coastal and inland locations, using rainfall values from neighbor stations as inputs, where the minimum, the first interquartile (Q1), the median, the third interquartile (Q3), the maximum, and the outlier values are represented. 193

Figure 5. 8. Seasonality performance of the different models (SVM, MLP, and RF) in all the stations and approaches 194

Figure 5. 9. **.** Scatter plot for predicted values in Torreblascopedro, using MLP, SVM, RF, and LI, during the different seasons 195

Figure 5. 10. Scatter plot for predicted values in Archidona, using MLP, SVM, RF, and LI, during the different seasons. 196

Figure 6. 1. Spatial distribution of Aroche, Conil, Córdoba, Málaga, and Tabernas in the Andalusia region, south of Spain……………………………………………………207

Figure 6. 2. Pearson correlation values of the assessed features in all the stations..... 209

Figure 6. 3. Mind map of the matrix data structure 211

Figure 6. 4. Forecasting approaches using configuration 1 as an example 212

Figure 6. 5. Methodology flowchart 213

Figure 6. 6. Single neuron architecture. I1, I2, I3, and I4 represent the inputs of the neuron, W1, W2, W3, and W4 correspond to the weights of every input, B is the bias, and O represents the output of the neuron after passing through an activation function 214

Figure 6. 7. Original Transformer architecture 216

Figure 6. 8. Sine / Cosine positional encoding for 31 days in a month (a) and 12 months in a year (b) 218

Figure 6. 9. The architecture of the proposed multi attention-based model 219

Figure 6. 10. Boxplot with RMSE values from all models and configurations in the different AWS, using 15 lag days (a) and 30 lag days (b) 234

Figure 6. 11. Boxplot with NSE values from all models and configurations in the different AWS, using 15 lag days (a) and 30 lag days (b) 235

Figure 6. 12. Boxplot with MBE values from all models and configurations in the different AWS, using 15 lag days (a) and 30 lag days (b) 236

Figure 6. 13. Scatter plot with the best NSE value for each model and location. 237

Figure 6. 14. Scatter plot with the best RMSE value for each model and location. 237

Figure 6. 15. Scatterplot with the best MBE value for each model and location. 238

Figure 6. 16. Scatter plot for measured vs. predicted values for (a) forecast horizon 1 in Tabernas, (b) forecast horizon 1 in Conil de la Frontera, (c) forecast horizon 7 in Tabernas and (d) forecast horizon 7 in Conil de la Frontera. 241

Figure 6. 17. Scatter plot for measured vs. predicted values for (a) forecast horizon 1 in Aroche, (b) forecast horizon 1 in Málaga, (c) forecast horizon 7 in Aroche and (d) forecast horizon 7 in Málaga. 242

Figure 6. 18. Scatter plot for measured vs. predicted values for for (a) forecast horizon 1 in Córdoba, (b) forecast horizon 7 in Córdoba 243

Figure 7. 1. Automated Weather Stations (AWS) in the Andalusia region and which region does each AWS belongs to RIA……………………………………………………..256

Figure 7. 2. Methodology 263

Figure 7. 3. Multilayer perceptron architecture for a model with two inputs ($x^{(0)}_1$ and $x^{(0)}_2$), one single hidden layer containing 3 neurons ($x^{(1)}_1$, $x^{(1)}_2$ and $x^{(1)}_3$) and one output (y1). W represents the weights for each connection. 265

Figure 7. 4. BO results for the ELM hyperparameters, where a) represents the chosen activation function and b) the number of neurons in the hidden layer. Rbf_l2 represents a

Radial Basis Function using L2 regularization, sigm represents a sigmoid activation function and tanh represents a hyperbolic tangent activation function 267

Figure 7. 5. BO results for the MLP hyperparameters, where a) represents the chosen activation function, b) the number of training epochs, c) the number of Hidden layers and d) the number of neurons on each hidden layer. Relu represents Rectified Linear Unit activation function 267

Figure 7. 6. RMSE values for the best configurations in region 1. See table 2 for further information about each location. 269

Figure 7. 7. RMSE values for the best configurations in region 2. See table 2 for further information about each location 269

Figure 7. 8. NSE values for the best configurations in region 1................................. 270

Figure 7. 9. NSE values for the best configurations in region 2................................. 270

Figure 7. 10. R^2 values for the best configurations in region 1. See table 2 for further information about each location. 271

Figure 7. 11. R^2 values for the best configurations in region 2. See table 2 for further information about each location. 271

Figure 7. 12. MBE values for the best configurations in region 1. See table 2 for further information about each location. 272

Figure 7. 13. MBE values for the best configurations in region 2. See table 2 for further information about each location. 272

Figure 7. 14. Mean monthly wind speed values for GRA02 (Puebla de Don Fadrique), SEV101 (IFAPA Centro de Las Torres.-Tomejil-Finca Tomejil), JAE08 (Úbeda) and COR02 (Adamuz)... 273

Figure 7. 15. RMSE (Root Mean Square Error) seasonal values for all the AWS using configuration 4... 278

Figure 7. 16. . NSE (Nash-Sutcliffe Efficiency model) seasonal values for all the AWS using configuration 4. 279

Figure 7. 17. R^2 (Coefficient of Determination) seasonal values for all the AWS using configuration 4... 279

Figure 7. 18. MBE (Mean Bias Error) seasonal values for all the AWS using configuration 4... 279

List of acronyms

AI	Artificial Intelligence
AgroML	Agronomy Machine Learning
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
AWS	Automated Weather Station
BESS	Battery Energy Storage System
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BO	Bayesian Optimization
BC	Bristow-Campbell
CatBoost	Gradient Boosting with Categorical Feature Support
CNN	Convolutional Neural Network
DeepFM	Deep Factorization Machine
DOY	Day of Year
DNN	Deep Neural Network
DT	Decision Tree
DTR	Diurnal Temperature Range
ELM	Extreme Learning Machine
FAO	Food and Agricultural Organization
FFA	Firefly Algorithms
FFMLP	feed-forward multilayer perceptron
GA	Genetic Algorithms
GBDT	Gradient Boosting Decision Tree
GEP	Gene Expression Programming
GPR	Gaussian Process Regression
GRNN	Generalized Regression Neural Network
GWO	Grey Wolf Optimizer
HS	Hagreaves Samani
HSMOD	Local-calibrated Hargreaves Samani
IoT	Internet of Things
IDWM	Inverse Distance Weighting Method
IPCC	Intergovernmental Panel on Climate Change
LightGBM	Light Gradient Boosting Method
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression
MBE	Mean Bias Error
MFO	Moth Flame Optimization
MI	Mutual Information
MOY	Month of the Year

MLP	Multilayer Perceptron
MTR	Monthly Temperature Range
NN	Neural Network
NWP	Numerical Weather Prediction
PC	Pearson Correlation
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
PV	Photovoltaic plants
R	Correlation coefficient
R^2	Coefficient of Determination
RAM	Random-Access Memory
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RCP	Representative Concentration Pathway
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RRMSE	Relative Root Mean Square Error
RTD	Resistance Temperature Detector
SLFN	Single hidden layer feedforward neural network
SVM	Support Vector Machine
SVR	Support Vector Regression
WCA	Water Cycle Optimization Algorithm
WOA	Whale Optimization Algorithm
WMO	World Meteorological Organization
WNN	Wavelet Neural Network
WR	Wavelet Regression
XGBoost	Extreme Gradient Boosting

List of symbols

γ	Psychrometric constant
Δ	Slope of the saturation vapor pressure
ξ	Deviation of values to the margins in SVM
ε	Margins in SVM
α_j	Coefficient for each model in GPI
σ	Standard deviation
σ_x	Standard deviation of the input dataset
σ_{meas}	Standard deviation of the measured values
σ_{pred}	Standard deviation of the predicted values
a	empirical coefficients
b	bias coefficient
c	empirical coefficients
C	penalization coefficient in SVM
DTR_M	Mean diurnal temperature range
DTR_N	Minimum diurnal temperature range
DTR_X	Maximum diurnal temperature range
e_a	Mean actual vapor pressure
e_s	Saturation vapor pressure
$Energy_T$	Integral of the temperature values of a day
EPm	Pan evapotranspiration
ET	Reference Evapotranspiration
ETa	Actual evapotranspiration
ET_0	Reference evapotranspiration
$ET_{0\ HS}$	Reference evapotranspiration calculated using HS
ET_C	Crop evapotranspiration under standard conditions
ETc adj	Crop evapotranspiration under non-standard conditions
f	activation function
G	Soil heat flux density at the soil surface
k(q)	scaling exponent
K_T	Clearness index
m	Length of the full dataset
m_{TE}	Length of the test dataset
m_{TR}	Length of the train dataset
m_V	Length of the validation dataset
meas	Measured values
$\overline{meas}$	Mean of the measured values
n	Number of sunlight hours
N	Maximum theoretical sunlight hours
pred	Predicted values
$\overline{pred}$	Mean of the predicted values

Ra	Extraterrestrial solar radiation
ReLU	Rectified Linear Unit
Rs	Shortwave radiation
Rso	Shortwave clear-sky radiation
Rn	Net radiation
Rnl	Net solar radiation
Rns	Net solar radiation
stats	Statistics values like (R^2, NSE, MSE and others)
tanh	Hyperbolic tangent
Tm	Mean daily air temperature
Tn	Minimum daily air temperature
T_X	Maximum daily air temperature
w	Weight coefficients
x	Inputs of a model
x^*	Inputs of a model standardized
$\bar{x}$	mean
y	Outputs of a model
u_2	Wind speed at 2 m above ground surface
u_z	Wind speed at z m above ground surface
z	height of measurement above ground surface

Chapter 1

Introduction

1. The current situation

During the last decades, the world's population is growing and their living standards are increasing remarkably rapidly. Therefore, it has become imperative to optimize the consumption of natural resources while not affecting the production of two main sectors, food, and energy, among others.

Taking into account the first sector, water resources management is crucial to reducing its consumption and increasing food production, considering not only the aforementioned problems but also climate change. The escalating frequency and intensity of droughts, floods, and other weather-related phenomena are expected to exacerbate the already challenging task of managing water resources [1–3], which is highly influential in arid and semiarid regions like Andalusia, where the scarce precipitations and high water demands have developed a structural water deficit. Besides, future temperature and precipitation predictions from the three RCP scenarios (RCP4.5, RCP6.0, and RCP8.5) do not highlight a promising future, with average changes in temperature up to +11ºC and in precipitation up to -30% in the worse scenario (RCP 8.5). In these terms, the introduction of automation in agriculture will help to contribute to raising productivity, improving product quality, and helping to achieve the Sustainable Development Goals by 2030 (no poverty and zero hunger) [4]. Thus, introducing motorized mechanization and other technologies such as sensors and robots relying on machine learning models would significantly impact all the phases of the cycle of an automation system (Figure 1. 1), especially focused on optimizing water resources management.

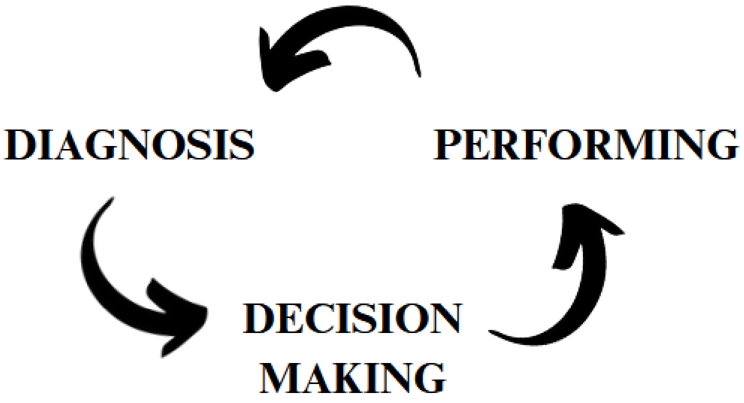


Figure 1. 1. The three-phase cycle of an automation system [4]

On the other hand, considering the energy sector, governments are investing in producing clean and sustainable energy, with the objective of minimizing the release of global warming emissions. As a result, the number of renewable energy plants is increasing by leaps and bounds, with special consideration for PV plants due to their easier and cheaper installations on both, high and low scales[5,6]. As an illustration, the global cumulative PV installations have been growing at an annual rate of 32.5% in the period 2010 to 2021, being the total cumulative installed PV capacity around 1.15TW in March 2022 [7,8]. Focusing on the European Union, the cumulated PV capacity reached 165 GW in 2021, with Germany, Spain, Poland, and the Netherlands being the top installing countries, each with more than 1 GW of installed capacity. Hence, developing accurate solar radiation models would be crucial for determining appropriate solar plant locations, for decision-taking criteria on BESS about whether to save the current energy production according to future availability and needs, and for agronomy, since solar radiation is highly important in processes like transpiration and evaporation.

Thereby, developing novel algorithms based on AI to outperform agrometeorological estimations and predictions variables such as solar radiation, precipitation, and reference evapotranspiration, among others, is impactful to contribute to the current society situation. Not only is high accuracy required, but reducing the number of agrometeorological measures needed is also important.

2. Agrometeorological variables

2.1.Air Temperature

The air temperature quantifies the amount of solar radiation absorbed by the atmosphere plus the heat emitted by the Earth. It is the most common agrometeorological measured variable because of its high accuracy even with low-cost sensors [9,10], as well as being very important in modeling other parameters.[11–13]. According to the IS, it is measured in K, although the unit ºC is very common due to its similarity to K (ºC = K−273).

The air temperature can be measured using different technologies. Firstly, one of the most common sensors is the Resistance Temperature Detector (RTD), which consists of a fine metal wire (usually platinum) whose electrical resistance changes with the temperature. The equivalences between resistance and temperature for a PT100 (100Ω at 0ºC) and a PT1000 (1000 Ω at 0ºC) are standardized in IEC 60751 [14]. One of the main advantages of RTDs is that they have high accuracy and stability, which allows their use in industrial environments. However, they have a high price and they require a power supply. Secondly, thermocouples are another type of sensor based on a different principle. In this case, the thermocouple is made of two different metallic wires joined at one end (this point is known as the hot junction), which is usually in contact with the object to be measured. The temperature at this spot induces a voltage at the other end of the wire (cold junction), being able to determine the first according to the second. The thermocouples have a lower price than RTDs, although they are less stable and less accurate. Thirdly, the thermistors have similar functionality to RTDs but they are composed of polymers or ceramics instead of metals. These sensors usually achieve a very high precision within a limited temperature range.

In Automated Weather Stations (AWS) the temperature must be measured at 2m height, following the World Meteorological Organization (WMO) standard, allowing its calculation on RTD, generally.

2.2.Relative humidity

The relative humidity is a measure of the water vapor density in the air compared to the maximum amount that can be held at a given temperature. It is expressed in a percentage, indicating how close the water vapor density is to being saturated, thereby, a 100% value means that the air cannot contain more moisture at that temperature. Additionally, the relative humidity is influenced by the temperature, warmer temperatures result in a higher moisture retention capacity and a decrease in relative humidity, and vice-versa.

In agronomy, this parameter is very important, since it can slow down the rate of evapotranspiration. Thus, a well-watered field located in a hot dry region will allow high evapotranspiration to happen, whereas well-watered fields in humid topical sites will decrease this rate because the air is closer to saturation [15]. Moreover, due to the high relationship between relative humidity and the probability of rain and/or the presence of clouds, it is also very important for the renewable energy sector.

In order to measure this variable, the hygrometer is used. It estimates relative humidity by measuring other parameters like temperature, pressure, and mass, among others, defining the type of device. Mechanical hygrometers, on the one hand, are based on the principle that organic substances expand/contract in response to humidity. Electrical hygrometers can be classified into two main types, the devices that measure the resistance or the capacitance variations according to the absorbed amount of water. The dew point hygrometer can calculate the relative humidity from the dew point temperature. Finally, the psychrometer consists of two thermometers, one of which has a wet wick around its bulb. The psychrometer works on the principle that when water evaporates from the wet wick, it cools the thermometer bulb, causing the temperature reading to be lower than the dry bulb thermometer. The difference in temperature readings between the two thermometers can be used to calculate relative humidity.

2.3.Solar radiation

Solar radiation refers to the electromagnetic radiation emitted by the sun that reaches the Earth. In the IS, the solar radiation is measured in W/m^2, although it is also common to find kWh/m^2, J/m^2, and MJ/m^{2}day as alternative units

It is the primary energy source for the climate and weather systems, as well as PV plants, among others. It plays a crucial role in agronomy and renewable energy sectors alike.

- **Extraterrestrial radiation**. It refers to the amount of solar radiation that reaches the outermost layer of the Earth's atmosphere, representing the amount of energy that the sun emits and reaches the atmosphere per unit area (on a horizontal surface). Its value can be theoretically calculated using the date and the latitude [16]. Figure 1. 2 shows the extraterrestrial radiation values for the northern and southern continental sites of Andalusia in MJ/m^{2}day.

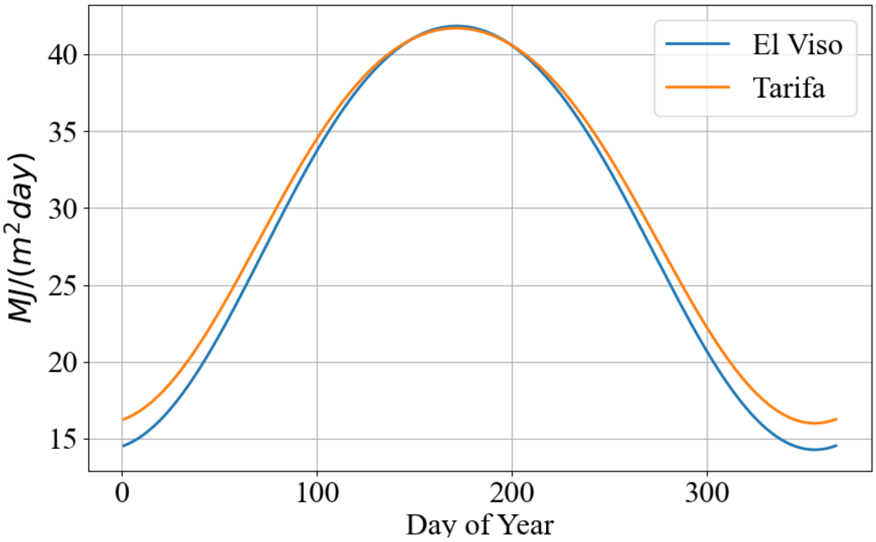


Figure 1. 2. Extraterrestrial solar radiation in the most northern and southern continental sites of Andalusia (El Viso and Tarifa, respectively)

- **Shortwave radiation**. As solar radiation enters the atmosphere, there is a portion that is absorbed, reflected, and scattered due to clouds and atmospheric gases. The amount of energy that eventually reaches a horizontal plane in the Earth's ground is called shortwave radiation. The amount of radiation in clear-sky conditions is called clear-sky shortwave radiation (Rso). Its value, as well as the clearness index (K_T), can be calculated using the methodology proposed by Allen [15] (equations 1.1 and 1.2).

$$R_{SO} = K_T \, Ra \quad \text{Eq. 1. 1}$$

$$K_T = 0.75 + 2 \cdot 10^{-5} \, z \quad \text{Eq. 1. 2}$$

The ratio between the shortwave radiation and the shortwave radiation in clear sky atmospheric conditions is known as the relative shortwave radiation, which is a way to determine how cloudy a day is. Their values range from 1 on a clear sky day to 0.33 on a dense cloudy day.

- **Relative sunshine duration**. Another important ratio in order to determine the cloudiness of the atmosphere is the ratio between the actual daylight hours (n) and the theoretical maximum possible duration (N). On clear-sky conditions, n should be equal to N.
- **Albedo**. The albedo(α) quantifies the amount of solar radiation reflected by the surface, which is sensitive to the characteristics of the surfaces and vegetation. The amount of solar radiation that does not reflect is called net solar radiation (R_{NS}).
- **Net longwave radiation**. The Earth absorbs solar radiation, which is converted into heat energy. However, part of this is lost by several processes, for example, emitting radiative energy. This radiation is known as net longwave radiation (Rnl). Part of this energy is reabsorbed by the atmosphere, whereas the rest is lost in space.
- **Net radiation**. The net radiation reflects the balance between Rns and Rnl, which takes into account the energy reflected, absorbed, and emitted by the Earth's surface. Rn is usually positive during a 24-hour period, positive during the day, and negative during the night.

Figure 1. 3 shows a visual representation of the different solar radiation components.

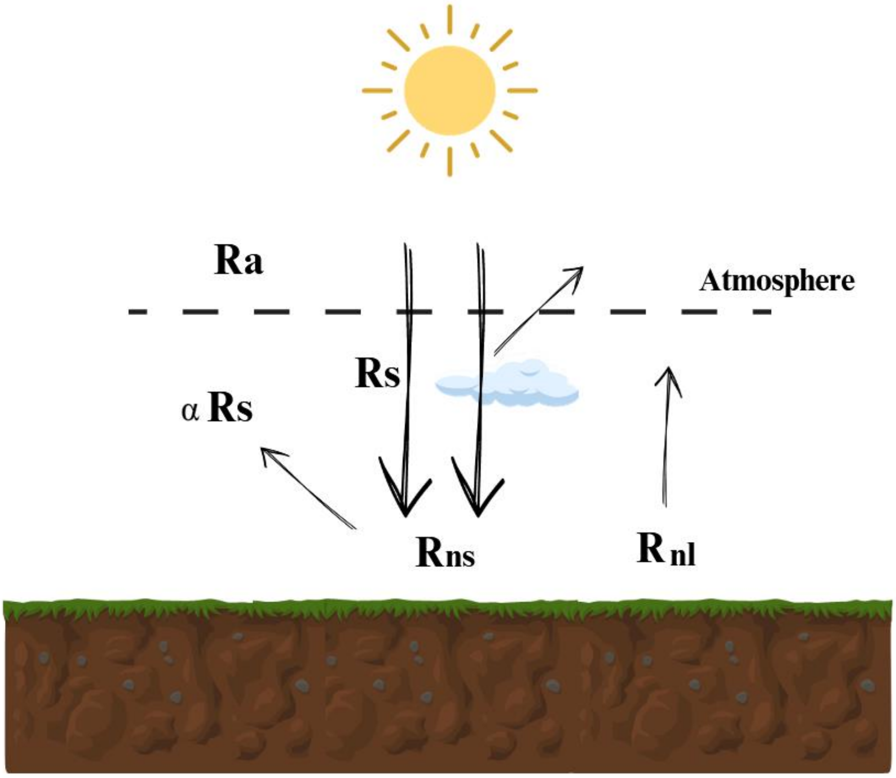


Figure 1. 3. Solar radiation components.

Solar radiation can be measured using three main devices, the pyranometers, the radiometers, and the solarimeters. The pyranometer is designed to measure the total amount of solar radiation that reaches a particular site, combining direct and diffuse solar radiation. Besides, the radiometer measures the amount of electromagnetic radiation presented in a particular environment, being able to detect wavelengths outside the visible light spectrum. Eventually, the solarimeter is specifically designed to quantify the intensity of sunlight. On the other hand, solar radiation can be estimated from n and N using the Angstrom formula [17], where n is measured using a Campbell-Stoke sunshine recorder.

2.4.Wind speed

The wind speed is composed of two main components, the velocity, which quantifies the intensity, and the direction, which refers to the wind's trajectory. According to the IS, the velocity is measured in $m \cdot s^{-1}$ using the anemometer. Besides, the direction is commonly measured by counting the time the wind blows in each direction quadrant (North, East, South, and West).

It is worth noting that wind speed has different values depending on the height due to the surface friction. The anemometer is standardized and located at 10m in meteorology AWSs and at 2 or 3m in agrometeorological AWS. The values can be estimated using Eq 1.3

$$u_2 = u_z \frac{4.87}{\ln(67.8 \cdot \mathrm{z} - 5.42)} \qquad \text{Eq. 1. 3}$$

2.5.Precipitation

The precipitation quantifies the amount of water that falls from the atmosphere to the surface. It is measured using a rain gauge, which automatically records a fixed water volume using a rocker arm system and buckets. The rain gauge measures the amount of rainfall by collecting water in a funnel, which then flows into one of two tipping buckets located at one end of a rocker arm. When one of the buckets fills up, it tips over and empties the water, while positioning the other bucket under the funnel to start its filling process. As each bucket tips over, the rocker arm briefly contacts a tongue for a few milliseconds. The amount of water collected is measured by counting the number of tips or pulses, with each pulse equivalent to a fixed rainfall measure.

The precipitation is measured in l, which is a unit of volume in IS, quantifying the amount of water in a given area. Besides, it can be measured in mm, which is a unit of length, corresponding to the height of the fallen rain in a 1 m^2 area. The equivalences are 1mm = 1 lm^{-2}.

2.6.Reference evapotranspiration

One of the most important agronomical parameters is evapotranspiration, which quantifies the actual water needs taking into account the crop characteristics, its management, and the weather parameters. The concept encompasses the water needs in crops that are produced through evaporation and transpiration. Evaporation is the process in which water is evaporated from a surface, whereas transpiration corresponds to the process in which the water contained in plant tissues is transferred to the atmosphere through the stomata, mainly. Both processes simultaneously happen without having an easy method to distinguish one another.

The evapotranspiration can be accurately determined using lysimeters. There exist different types of devices according to their functionality basis such as weighing, drainage, and wick lysimeters, among others, although the weighing lysimeter is most commonly used. It consists of a large buried container having plants and the soil they grow in. It is equipped with a system of load cells and sensors that allow measuring the weight of the entire system (plants, soil, and lysimeter) with high precision. By regularly weighing the lysimeter over time, it can be determined how much water is being lost from the system due to ET.

The main factor affecting evapotranspiration is the weather condition in the first term, being the air temperature, relative humidity, solar radiation, precipitation, and wind speed, among others, the principals. Besides, the crop characteristics such as the crop type and its development stage have to be considered as well. Eventually, the management conditions as the presence of fertilizers and well-watered irrigations also affect evapotranspiration. According to these factors, three main evapotranspiration variables can be derived: (I) the reference evapotranspiration (ET_0), which measure the water needs of a reference surface (standardized characteristics like the albedo and the crop height, among others) and management conditions, only affected by the weather parameters; (2) the crop evapotranspiration under standard conditions (ET_C), which considers the crop features in the formula; and (3) the crop evapotranspiration under non-standard conditions (ETc adj), considering the actual crop characteristics, management, and environmental conditions.

Since the cost associated to measure, ETc adj is very high, the use of ET_0 is considered an alternative for making water resources management decisions. The reference surface in this variable is a hypothetical grass reference crop assuming a 0.12m height, an albedo of 0.23, and a surface resistance of $70 sm^{-1,}$ and. It can be mathematically expressed using the FAO PM equation [15] as Eq 1.4:

$$\mathrm{ET_0} = \frac{0.408\Delta(\mathrm{R_N} - \mathrm{G}) + \gamma \frac{900}{\mathrm{T}_m + 273} \mathrm{u_2}(e_s - \mathrm{e}_a)}{\Delta + \gamma(1 + 0.34u_2)} \qquad \text{Eq. 1. 4}$$

where 0.408 corresponds to a coefficient ($MJ^{-1}\ m^2$ mm), Δ is the slope of the saturation vapor pressure versus temperature curve (kPa $°C^{-1}$), R_N is the net radiation calculated at the crop surface (MJ m^{-2} day^{-1}), G is the soil heat flux density at the soil surface (MJ m^{-2} day^{-1}), γ is the psychrometric constant (kPa $°C^{-1}$), Tm is the mean daily air temperature (°C), u_2 is the mean daily wind speed at 2 m height (m s^{-1}), and e_s and e_a are the saturation vapor pressure and the mean actual vapor pressure, respectively (kPa). The Eq 1.4 to determine ET$_0$ has been widely applied in different climate conditions [18–21] and countries, becoming the sole standard methodology for FAO.

However, the geographical density of AWS measuring all the meteorological measurements for FAO PM is very low, especially in rural areas and developing countries [22] due to the high cost associated with installation, maintenance, and calibration. Thus, novel models had been developed using a limited dataset, like the Hargreaves Samani [23] or the Bristow Campbell [24] equations which determine solar radiation and reference evapotranspiration using the maximum and minimum air temperature and the extraterrestrial solar radiation, which had been assessed in different countries and climate regions like Italy, Switzerland, Turkey, China, Mexico, and Spain [25–30], among others.

3. Automated Weather Stations in Andalusia

In Andalusia (Southern Spain), there coexist three main AWS networks, the Agroclimatic Information Network (RIA), the State Meteorological Agency (AEMET), and the Crop Health Alert and Information Network (RAIF). Figure 1. 4 show the geographical location of all the AWS from the RIA, RAIF, and AEMET networks.

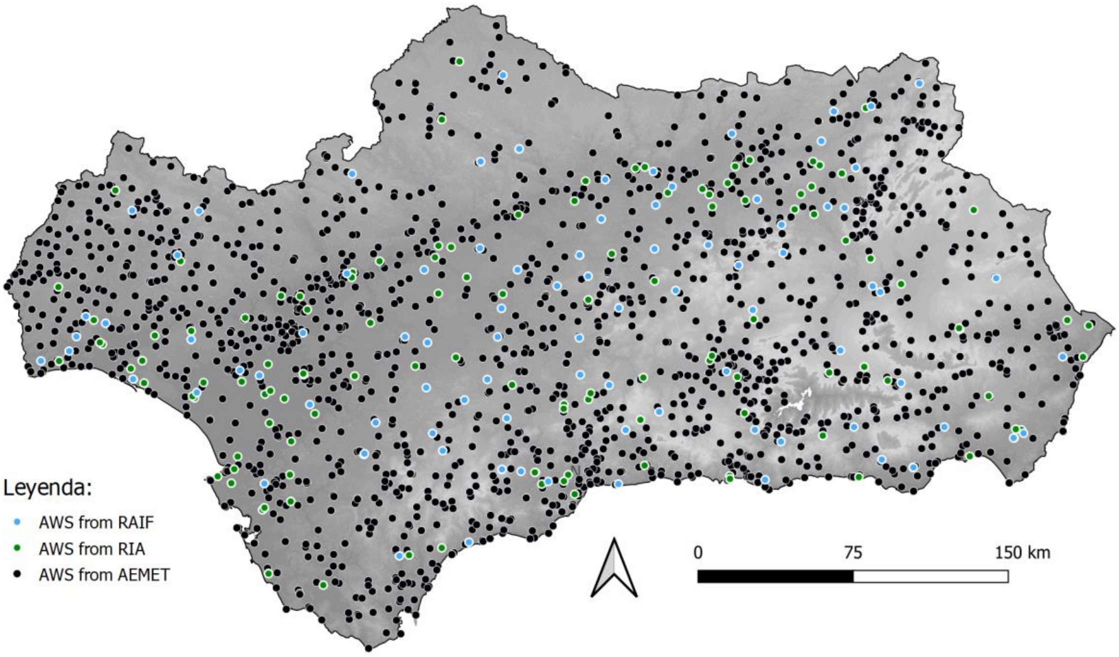


Figure 1. 4. AWS sites from AEMET, RIA, and RAIF

The RIA AWS network consists of 122 sites measuring wind speed and direction at 2 meters above ground level, solar radiation, relative humidity, air temperature, and precipitation using the sensors in Table 1. 1. Sensors specifications from the RIA AWS network [31]. The dataset covers a period from December 1999 to June 2022 with daily sampling frequency, which can be freely downloaded at the following link: https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web/ (Accessed on March 17th, 2023). Besides, the RAIF AWS network consists of a total of 79 AWS also measuring air temperature, relative humidity, solar radiation, accumulated precipitation, and the wind speed and direction at 10 m height. The dataset covers the period from

January 2002 to December 2021, which can be freely downloaded at the following link: https://www.juntadeandalucia.es/datosabiertos/portal/dataset/raif-clima (Accessed on March 17th, 2023). It is worth noting that both RIA and RAIF networks offer free access to daily records, but hourly records are only available upon request and for research purposes. Moreover, the AEMET AWS network is composed of a total of 1754 AWS measuring the air temperature and/or precipitation, thus they are limited compared to RIA and RAIF. Furthermore, daily records are available upon request, but only for research purposes and they are subject to payment.

Table 1. 1. Sensors specifications from the RIA AWS network [31]

Parameter	Sensor	Range	Accuracy
Solar radiation	Skye SP1110	350 to 1100 nm	±5 %
Relative Humidity	Humicap 180	0.8 to 100 %	±3 % (90-100%) ±2 % (0.8-90%)
Wind speed	Young 05103	0 to 60 m s^{-1}	±0.3 m s^{-1}
Wind direction		0 to 360°	±3 °
Temperature	PT1000	-39.2 to 60 °C	±0.3 °C
Precipitation	ARG 100		0.2 mm/tip

4. Quality Control Procedures

Ensuring the use of reliable, high-quality data is the first requirement when conducting studies using time series datasets and modeling tasks. Thereby, carrying out quality control procedures is a critical process in data science to ensure the accuracy and reliability of analysis results. It involves evaluating data quality in order to identify and correct errors and/or inconsistencies, whose problems can be associated with different sources:

- **Problems associated with sensor failures**. Like any electronic component, sensors have a lifespan, and components can break. There also exists the possibility of failure associated with a blow or adverse weather conditions. This would generate the storage of a null or incorrect value.
- **Problems associated with sensor drifts**. Sensor drifts are a common problem in time series datasets, they can deviate from their original calibration due to factors such as temperature variation, vibration, or prolonged use. When this happens, the data measured by the sensors can be incorrect and not accurately represent the reality being measured. This can be a serious problem, as inaccurate data can lead to incorrect conclusions and decisions. There are various techniques to address sensor drifts, such as recalibrating the sensors and applying correction models to adjust the measured data. Additionally, it is important to regularly monitor sensors and take preventive measures to minimize deviation from their original calibration.
- **Problems associated with datalogger failure**. A data logger is a device used to automatically collect data over time. These devices can be very useful for collecting data in remote environments and in real time. However, a common

problem is data loss due to device failure. If a datalogger breaks before the data has been collected/downloaded, all the stored information contained on the device will be lost. To minimize the risks associated with datalogger failure, it is important to use high-quality devices and ensure regular maintenance and checks are performed. Additionally, it is important to have a contingency plan in case the device fails or breaks. This could include the use of redundant devices and/or manual data collection, among others.

- **External problems**. Studies such as Estévez et al [36] detect the occurrence of spurious signals in precipitation measurements caused by irrigation systems.

The different methodologies to identify and flag erroneous data can be classified into range, step, internal consistency, persistence, and spatial consistency tests [31], for the different meteorological variables and the sampling rate. The **range tests** assess the sensor value considering the technical specifications and the climate limits (sensitive to the AWS site and the season). For example, temperature records outside the sensor range, or solar radiation measures above Rso, must be flagged as erroneous data. The **step tests** evaluate changes in values of the different variables at different frequencies, evaluating the excessive change rates of consecutive records. For example, in hourly solar radiation records, Meek and Hatfield [32] proposed that the difference between two consecutive records must be between [0, 555] Wm^{-2}. **Internal consistency tests** determine the quality of measures assessing variables recorded at the same location and time; two variables measured at the same site and time must be consistent, being flagged as erroneous otherwise. For example, T_X must be higher than Tm and Tn on any frequency basis. The **persistence tests** evaluated the variability of records, since the sensor state a constant value when failing. Among other statistics, the standard deviation is observed not to be below an acceptable minimum, since constant values have a zero standard deviation value. For example, temperature values (T_X, Tm, and Tn) cannot be the same for three consecutive days. Finally, **spatial consistency tests** compare data between the surrounding weather stations and flag suspects records whose difference exceeds a limit. It is worth noting that the observations are weighted regarding the distance between them, although, due to the weather variability, these tests are commonly used in temperature records only.

5. Feature engineering and selection

Feature engineering corresponds to the process of studying the raw data (time series datasets from an AWS, in this case) and transforming it into different and/or novel features to outperform estimations and predictions' modelings. The main goal during this process is to extract meaningful information from the given dataset so that models using these new features outperform those using only the raw data.

Several techniques can be used for feature engineering, and the chosen technique depends on the nature of the data and the problem being solved. One common technique is to use domain knowledge to engineer features that are known to be relevant to the problem. For example, instead of using the coordinates, a better alternative is to calculate extraterrestrial solar radiation, which gives more valuable information. In these terms, temperature-based features had been determined such as the time when T_X occurs, the time when Tn occurs, and $Energy_T$ (the integral of semi-hourly temperature records in a day), among others. Figure 1. 5 shows the temperature records of 24th May 2015 at the AWS from RIA located in Cordoba

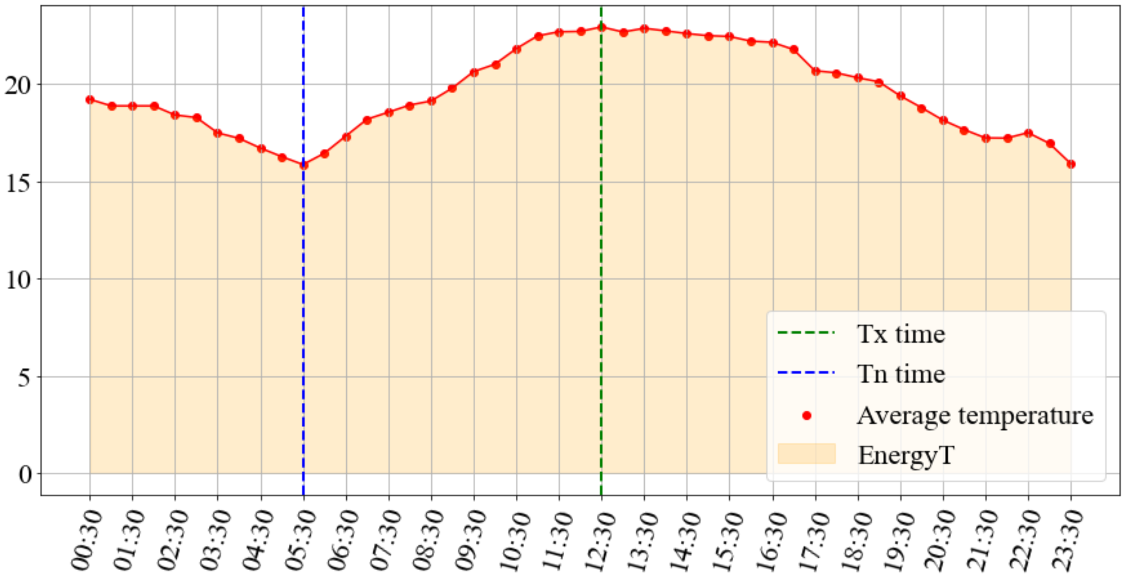


Figure 1. 5. Semi-hourly temperature and relative humidity records on 24th May 2015 at the AWS from RIA located in Cordoba. The Tx time and Tn time represent the time at which T_X and Tn occur, respectively

Another technique is the use of statistical methods to find the features or new ones that are more correlated with the target variable, for example, the use of Principal Component Analysis (PCA). PCA is commonly utilized to identify patterns and extract information from the original dataset, which is usually used for dimensionality reduction purposes [33]. Its main goal is to transform the original dataset into a new coordinate system, using new features calculated as a linear combination of the originals (known as principal components).

The **feature selection** process refers to the selection of the best feature subset that obtains the most accurate modeling. It is a crucial phase in data analysis and machine learning applications because getting irrelevant variables leads to a decrease in performance, using too many features significantly increase the computation time and may leads to overfitting, and modeling with very few input variables can result in underfitting performance. Usually, two main approaches are considered, the trial and error technique, where different and even all the possibilities are tested, and two main statistical filtering methods, the Pearson Correlation (PC) and the Mutual Information (MI) coefficients. Pearson Correlation measures the linear relationship between two variables, ranging from -1 to +1, where values close to 0 indicate the absence of correlation, and values close to +1 and -1 refer to a high positive and negative linear correlation, respectively. On the other hand, MI measures the information contained in one variable about the other, ranging from 0 to + infinity. Values close to 0 indicate the absence of information while higher values refer to stronger relationships. Both PC and MI have been evaluated for agrometeorological works [34–36]

6. Machine learning models

6.1.Linear Regression

Linear regression is one of the most basic and widely chosen models in supervised learning as a baseline due to both its simplicity and low computational cost. It is a statistical technique used to find a linear relationship between an input variable (independent) and an output variable (dependent) in a dataset. In other words, linear regression predicts the output value from linear combinations of the inputs. The goal of linear regression is to find a straight line that best fits the data so that it can be used to predict future values. Depending on whether the model has one or more inputs it can be classified into simple linear regression (one single input predicts one single output) and multiple linear regression (multiple inputs predict a single output).

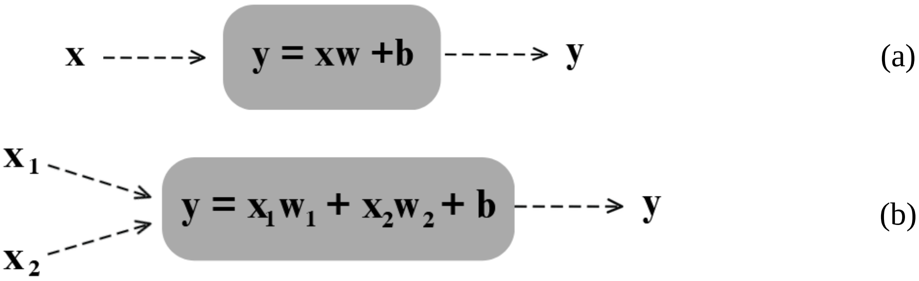


Figure 1. 6. Example of linear regression models: a) simple linear regression and b) multiple linear regression

The weights of linear regression models are calculated using the least square method, which consists of finding the weights and bias values that minimize the sum of squared differences between the inputs and output. The process is carried out iteratively during the training stage until the fittest set is found (obtaining the minimum error).

6.2.Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) models are inspired by the complex structure and functionality of neurons in the human brain, where a net of interconnected neurons process the information given by the senses (input features) and determine the most appropriate stimuli (output estimation/prediction). They were first introduced by Rosenblatt back in 1958 [37], although the actual capabilities required the development of higher computational power that came in the later decades.

Its structure is composed of multiple layers of interconnected nodes (also known as neurons) that works very similarly to multiple linear regression in a low-level view and can capture complex patterns between the inputs and outputs variables, in a high-level view. The different layers can be classified as input, hidden, and output layers, referring to the inputs, the neurons architecture, and the outputs of the model, respectively. Figure 1. 7 shows an example of an MLP model having 4 inputs, 1 output, and 3 hidden layers with 5, 4, and 3 neurons, respectively. Besides, Figure 1. 8 shows the architecture of a single neuron architecture.

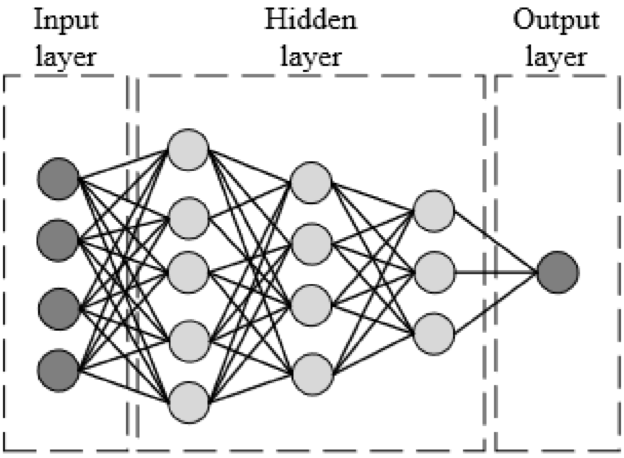


Figure 1. 7. Multilayer Perceptron architecture example for a model having 4 input features, 1 output, and 3 hidden layers with 5, 4, and 3 neurons, respectively

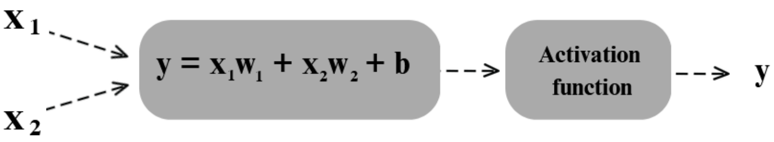
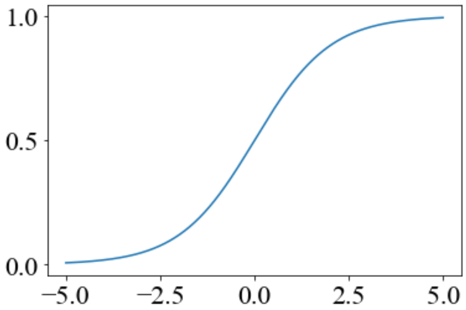


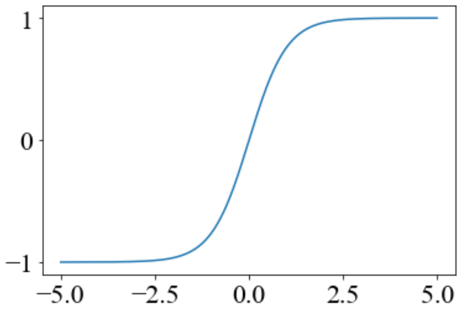
Figure 1. 8. Architecture of a single neuron of an MLP model

The activation function is a mathematical function that introduces non-linearity behaviors at the output of each neuron, allowing MLP to learn and deploy more complex relationships than linear regression. There are different types of activation functions, the sigmoid function, the hyperbolic tangent (tanh), and the rectified linear unit function (ReLU). Figure 1. 9 shows the mathematical equation and its plot.

a) Sigmoid $f(x) = \frac{1}{1 + e^{-x}}$



b) Tanh $f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$



c) ReLU $f(x) = \begin{cases} 0 \; if \; x < 0 \\ x \end{cases}$

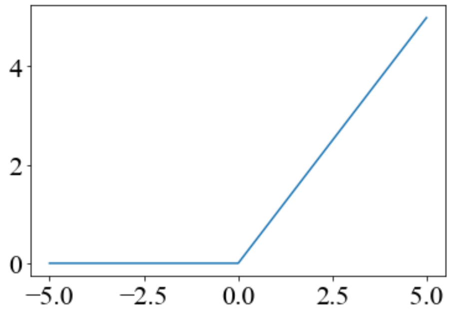


Figure 1. 9. Activation function: a) sigmoid, b) ReLU, and c) tanh

Another important aspect of MLP is the process of training, where all the weights and biases are automatically adjusted, known as backpropagation. This algorithm is carried out iteratively using an optimization function that calculates the gradient of the error according to each weight and bias, telling the algorithm how much to adjust in order to minimize the modeling error. This calculation is repeated in many iterations, known as an epoch, whose value depends on the complexity of both the architecture and the process to model; the higher the complexity, the higher the number of epochs. A bad choice in this aspect may result in overfitting or underfitting.

MLP models have been proven to successfully work in agrometeorological applications, being one of the most chosen algorithms due to both, their good general efficiency and their conceptual simplicity. For example, for estimating/predicting solar radiation [38–40], precipitation [11,41–44], and reference evapotranspiration [45–51], among others. On the other hand, the computational cost of deep architecture (also when a very large dataset is studied) during the training is very high.

6.3.Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) models correspond to a single hidden layer feedforward neural network (SLFN) architecture with the main difference that the input weights are automatically set with a value ranging from 0 to 1, and the output weights are analytically calculated minimizing the minimum norm least squares solution (like in a general linear system). The absence of a backpropagation training process makes these models very fast to adjust, although a very high RAM is required for very long datasets.

They were first introduced by Huang [52], inspired by the promising learning capabilities of SLFN using almost any non-linear activation function [53]. Research indicates that ELM can estimate and predict agrometeorological variables with high precision [54–64].

6.4.Support Vector Machine (SVM)

Support Vector Machine (SVM) for regression tasks, also known by some researchers as Support Vector Regression (SVR) has a similar conceptual basis to linear regression, where the fittest hyperplane is determined in order to model a system, but introducing an extra complexity layer where the acceptable error can be defined (slack variables). The idea behind slack variables is to allow some data points not to be contained in the defined hyperplane but to penalize them in the cost function. During the training stage, the algorithm must minimize the sum of the slack variables.

The introduction of slack variables benefits in determining hyperplanes that generalize better in real-world applications with unseen data, minimizing the possibility to overfit. Figure 1. 10 shows an illustrative example of slack variables in SVR models. Besides, the optimization algorithm is mathematically expressed as Eq 1.5 constrained by Eq 1.6.

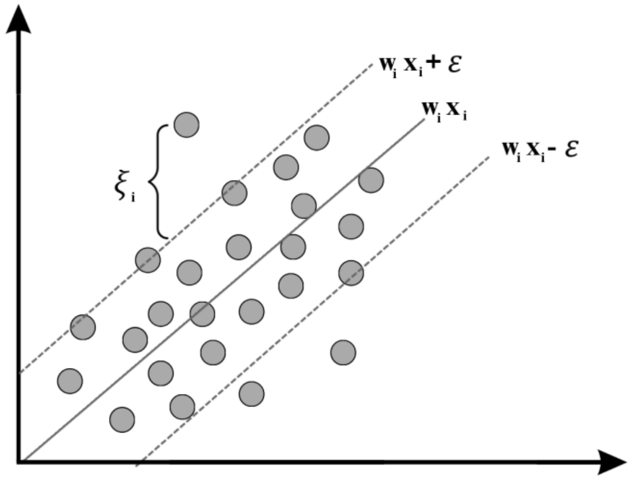


Figure 1. 10. Illustrative example of slack variables in SVR models

$$MIN\left(\frac{1}{2}\,\|w\|^2 + C\,\sum_{i=1}^{m_{TR}}|\xi_i|\right) \geq 0 \qquad \text{Eq. 1. 5}$$

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i| \qquad \text{Eq. 1. 6}$$

where w_i corresponds to the weight vector, x_i to the input vector, y_i to the output vector, ε represents the margins, ξ represents the deviation of values to the margins, C is a coefficient to penalize deviation to the margins, and m_{TR} the length of the training dataset.

An important characteristic of SVM is the use of kernel functions, which enable the models to operate in a high-dimensional feature space. The definition of kernel functions is given by Vapnik [65]: "*Given a mapping function Ø:X→V, we call the function K:X→R defined by K(x, x') = (Ø(x), Ø(x'))v, where (·, ·)v denotes an inner product in V, a kernel function*". The primary benefit of using kernel functions in SVM is that they allow them to effectively find complex and non-linear relationships between input features and the output, by transforming the data into a higher dimensional space. Besides, kernel functions easy SVM to work with large datasets without compromising the computing cost. The most typical kernel functions are the linear, polynomial, and radial basis function (RBF) kernels.

SVM models have been assessed at different sectors [66–69], and agrometeorological variables [70–77] obtaining accurate results even with default hyperparameter configurations [78].

6.5.Random Forest (RF)

Random Forests (RF) are based on Decision Tree (DT) models. DT models work by iteratively making decision rules that help to minimize the uncertainty in the output prediction. During the training stage, the model finds which feature and value range must be consulted in each decision. A useful analogy to describe this model is the classic game "Guess Who?" In this game, players ask yes-or-no questions to identify the opponent's selected character, such as "Is it male or female?" or "Does the character wear glasses?" Each of these questions is equivalent to a decision rule. Thus, finding the right question and its order is crucial, which is the main goal of DT models during the training phase.

The main advantages of these kinds of models are: (1) that they are easy to interpret since all decision rules can be graphically plotted; (2) they can handle missing data during the training set; (3) they can work with different data types with no preprocessing requirements (normalization and/or standardization); (4) decision trees can perform automatic feature selection, eliminating the less important features to enhance the model accuracy and reduce its training time; and (V) these models are non-parametric, meaning they make no assumptions about the underlying data distribution or the relationship between the input and output data (unlike models such as linear regression, based on the assumption of a linear relationship). This makes the model more flexible and capable of adjusting to a wide variety of applications.

In these terms, RF is a type of machine learning model that is based on creating a large number of independent decision trees, known as random decision trees, in addition to combining their results. In this model, each tree is trained with a random sample of the dataset, giving each of it a different perspective on the problem in order to obtain a better performance. These types of models that rely on predicting a value using the combination of different models' predictions are known as ensemble modeling.

RF models have similar advantages to DT models, such as the ability to handle large and complex datasets, and the ability to handle a large number of features. In addition, because the final prediction is the result of the combination of different models, they are less prone to making specific errors under specific conditions, since an erroneous value is corrected by the rest of the predictions. On the other hand, the computational cost is higher because several models have to be trained, although this computational load can be distributed in different computers, greatly reducing the computational time.

The process of building a random forest consists of two main phases: the creation of decision trees and the combination of their results:

1. **Creation of random decision trees**: Each tree is built from a random subset of training data and a random subset of features, greatly increasing diversity in modeling. Just as for tackling a project or business, the vision of different types

of professionals and viewpoints is beneficial, using multiple models trained on different features and subsets of data minimizes the final prediction error [79–82].

2. **Combination of results**: Once all trees are created, their results are combined to form a final prediction. Different methodologies are used in this phase, depending on the system to be modeled. In general, for classification, the majority vote result is used, where each tree votes for a class, and the class with the most votes is selected as the final prediction. In regression, the mean or median of the predictions is commonly applied. Additionally, for more complex modeling, it is common to find machine learning models that generate a final prediction from these individual predictions [83].

7. Hyperparameter selection

Hyperparameter selection is crucial in the process of building machine learning models. It consists of selecting the optimal hyperparameter set that makes the model yield the best performance for a given dataset. Choosing the correct hyperparameter set can modify the performance from bad to outstanding.

Various methodologies can be used to achieve this goal. One of the most common, albeit the least reliable, is the trial-and-error technique [77,84], which involves evaluating a set of values for each hyperparameter and selecting the best set based on performance. However, it should be noted that this approach does not provide any guarantee that the selected hyperparameter set is the most appropriate one for the given problem. A better solution would be to automatically carry out this methodology in a loop, such as Grid and Random search. The main difference is that Grid search evaluates all the possible values from the hyperparameter space (similar to a brute force algorithm), whereas Random search assesses random values. Some authors have evaluated that Random search performed the same as Grid search but with a lower computational cost [85,86]. Finally, the best solution is to carry out an optimization algorithm to solve this problem, for example, Bayesian Optimization (BO), Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Whale Optimization Algorithm (WOA), and Genetic Algorithms (GA), among others. There are several algorithms inspired by different animal behaviors: (1) PSO is inspired by the behavior of bird flocking or fish schooling, where all the particles move through the search space according to the best positions of other particles in the swarm; (2) GWO is inspired by the hunting behavior of grey wolves, where the population is divided into alpha, beta and delta wolves representing the best, second-best and third-best solutions; (3) WOA is inspired by the hunting behavior of humpback whales, where the whales move through the search space considering different strategies such as encircling the prey, bubble-net attacking and searching for prey. Additionally, BO has gained popularity recently due to its similarity to human behavior while carrying out this task [87,88] and its use in popular automated machine learning such as Auto-Keras, Auto-Sklearn, and Auto-Weka 2.0 [89] This algorithm is based on the Bayes theorem where a new set of hyperparameters is chosen and evaluated regarding previous results, alternating between two main strategies/functions (acquisition function), focusing on those areas with high uncertainty or those hyperparameter values believed to bring promising results.

8. AgroML

During this thesis, an easy Python tool for developing models for agrometeorological estimations and forecasts has been developed. It provides an abstraction layer to different machine learning libraries such as Tensorflow, Scikit-Learn, and Hpelm in order to make machine learning applications easier to develop and compare. Besides, different utils can be found, like data normalization/standardization, hyperparameter tuning, machine learning models like MLP, SVM, RF, ELM, XGBoost, and others, and agrometeorological calculations such as ET_0 and solar radiation using different methodologies.

The library can be freely found and used at https://github.com/Smarity/AgroML (Accessed 19th March 2023), previously used in recent works [46,47].

9. Hypothesis

The different hypotheses evaluated are:

A) Agrometeorological variables derived from temperature-based features can be accurately estimated/predicted using machine learning models, such as precipitation (works 1 and 4), solar radiation (work 3), and reference evapotranspiration (works 2, 5, and 6).
B) The use of better algorithms to improve hyperparameter selection such as BO can outperform the efficiency of models (work 2).
C) Clustering weather stations based on DTR can improve the accuracy of temperature-based agrometeorological variable predictions (work 6).
D) Multifractal properties of DTR can be used as a basis for developing a regional machine learning method for agrometeorological variable predictions (work 6).
E) The creation of regional models to deploy agrometeorological modeling only using temperature records in the whole process (works 3 and 6).
F) The incorporation of new temperature-based features such as $Energy_T$, $Hourmin_{Tx}$, and $Hourmin_{Tn}$ can improve the accuracy of agrometeorological variable predictions (works 2, 3, and 5)
G) Machine learning models such as MLP, SVM, and RF can outperform traditional temperature-based methods in predicting agrometeorological variables (work 2 and 3).
H) The accuracy of agrometeorological variable predictions varies depending on the geo-climatic conditions of the location being studied (work 6)
I) The aridity index and distance to the sea are factors that can affect the accuracy of agrometeorological variable predictions (works 2, 3, and 6)
J) Agrometeorological variables can be accurately predicted using temperature-based features even in areas with low rainfall or limited water resources (work 4).

10. Objectives

The main aim of this Ph.D. thesis is to research smart models to improve agrometeorological estimations and predictions considering two main statements: (1) the use of fewer measurements as input features, mainly focused on temperature-based datasets, and (2) to enhance the efficiency of models. Several secondary objectives have been proposed to achieve this purpose:

A) To develop accurate machine learning models for agrometeorological forecasting using temperature-based features and time series data (work 1 and 5).
B) To develop accurate machine learning models for agrometeorological estimations using temperature-based features and time series data (work 2, 3, 4, and 6).
C) To compare the performance of different machine learning models for agrometeorological forecasting in different geo-climatic conditions (work 5).
D) To evaluate the impact of new temperature-based features, such as $Energy_T$, hourminTx, and hourminTn, on the accuracy of agrometeorological forecasting models (works 2, 3, and 5).
E) To develop a clustering method based on multifractal properties of DTR for accurate weather station grouping and agrometeorological forecasting (work 6).
F) To assess the accuracy of temperature-based reference evapotranspiration (ET_0) estimations using machine learning models and compare them with the performance of the HS method (work 2, 5, and 6).
G) To investigate the seasonal variation in the accuracy of agrometeorological forecasting models and identify the most accurate models for each season. (works 2, 3, and 5).
H) To develop an open-source repository for agrometeorological forecasting models, which can be used in different geographic locations and climatic conditions (work 5).
I) To assess the impact of aridity index and distance to the sea on the accuracy of agrometeorological forecasting models in different regions (works 2, 3 and 5).

11.References

[1] 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. Cambirdge, UK and New York, NY, USA: 2022.

[2] Hervás-Gámez C, Delgado-Ramos F. Drought Management Planning Policy: From Europe to Spain. Sustainability 2019, Vol 11, Page 1862 2019;11:1862. https://doi.org/10.3390/SU11071862.

[3] Haile GG, Tang Q, Hosseini-Moghari SM, Liu X, Gebremicael TG, Leng G, et al. Projected Impacts of Climate Change on Drought Patterns Over East Africa. Earths Future 2020;8:e2020EF001502. https://doi.org/10.1029/2020EF001502.

[4] FAO, IFAD, UNICEF, WFP, WHO. The State of Food Security and Nutrition in the World 2022. Repurposing food and agricultural policies to make healthy diets more affordable. Rome: 2022.

[5] Jager-Waldau A, Adinolfi G, Batlle A, Braun M, Bucher C, Detollenaere A, et al. Electricity produced from photovoltaic systems in apartment buildings and self-consumption : Comparison of the situation in various IEA PVPS countries. Conference Record of the IEEE Photovoltaic Specialists Conference 2019:1701–10. https://doi.org/10.1109/PVSC40753.2019.8980484.

[6] Singh GK. Solar power generation by PV (photovoltaic) technology: A review. Energy 2013;53:1–13. https://doi.org/10.1016/J.ENERGY.2013.02.057.

[7] Jäger-Waldau A. Snapshot of photovoltaics – February 2022. EPJ Photovoltaics 2022;13. https://doi.org/10.1051/EPJPV/2022010.

[8] Chatzipanagi A, Jaeger-Waldau A, Cleret DLC, Letout S, Latunussa C, Mountraki A, et al. Clean Energy Technology Observatory: Photovoltaics in the European Union – 2022 Status Report on Technology Development, Trends, Value Chains and Markets. Luxrmbourg: 2022. https://doi.org/10.2760/812610.

[9] Strigaro D, Cannata M, Antonovic M. Boosting a Weather Monitoring System in Low Income Economies Using Open and Non-Conventional Systems: Data Quality Analysis. Sensors 2019, Vol 19, Page 1185 2019;19:1185. https://doi.org/10.3390/S19051185.

[10] Nsabagwa M, Byamukama M, Kondela E, Otim JS. Towards a robust and affordable Automatic Weather Station. Dev Eng 2019;4:100040. https://doi.org/10.1016/J.DEVENG.2018.100040.

[11] Estévez J, Bellido-Jiménez JA, Liu X, García-Marín AP. Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment. Water (Switzerland) 2020;12. https://doi.org/10.3390/w12071909.

[12] Sanikhani H, Kisi O, Maroufpoor E, Yaseen ZM. Temperature-based modeling of reference evapotranspiration using several artificial intelligence models: application of different modeling scenarios. Theor Appl Climatol 2019;135:449–62. https://doi.org/10.1007/S00704-018-2390-Z/FIGURES/8.

[13] Abraha MG, Savage MJ. Comparison of estimates of daily solar radiation from air temperature range for application in crop simulations. Agric For Meteorol 2008;148:401–16. https://doi.org/https://doi.org/10.1016/j.agrformet.2007.10.001.

[14] IEC 60751. Industrial platinum resistance thermometers and platinum temperature sensors. International Electrotechnical Commission 2022.

[15] Allen RG, Pereira LS, Raes D, Smith M. Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56. FAO - Food and Agriculture Organization of the United Nations; 1998.

[16] Stein JS, Hansen CW, Reno MJ. Global horizontal irradiance clear sky models : implementation and analysis. 2012. https://doi.org/10.2172/1039404.

[17] Angstrom A. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Quarterly Journal of the Royal Meteorological Society 1924;50:121–6. https://doi.org/10.1002/QJ.49705021008.

[18] Estévez J, Gavilán P, Berengena J. Sensitivity analysis of a Penman–Monteith type equation to estimate reference evapotranspiration in southern Spain. Hydrol Process 2009;23:3342–53. https://doi.org/10.1002/HYP.7439.

[19] Gavilán P, Estévez J, Berengena J. Comparison of Standardized Reference Evapotranspiration Equations in Southern Spain. Journal of Irrigation and Drainage Engineering 2008;134:1–12. https://doi.org/10.1061/(ASCE)0733-9437(2008)134:1(1).

[20] Martí P, González-Altozano P, López-Urrea R, Mancha LA, Shiri J. Modeling reference evapotranspiration with calculated targets. Assessment and implications. Agric Water Manag 2015;149:81–90. https://doi.org/10.1016/J.AGWAT.2014.10.028.

[21] Shiri J. Evaluation of FAO56-PM, empirical, semi-empirical and gene expression programming approaches for estimating daily reference evapotranspiration in hyper-arid regions of Iran. Agric Water Manag 2017;188:101–14. https://doi.org/10.1016/J.AGWAT.2017.04.009.

[22] Estévez J, Padilla FL, Gavilán P. Evaluation and Regional Calibration of Solar Radiation Prediction Models in Southern Spain. Journal of Irrigation and Drainage Engineering 2012;138:868–79.

[23] Hargreaves GH, Samani ZA. Estimating Potential Evapotranspiration. Journal of the Irrigation and Drainage Division 1982;108:225–30. https://doi.org/10.1061/JRCEA4.0001390.

[24] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric For Meteorol 1984;31:159–66. https://doi.org/10.1016/0168-1923(84)90017-0.

[25] Ravazzani G, Corbari C, Morella S, Gianoli P, Mancini M. Modified Hargreaves-Samani Equation for the Assessment of Reference Evapotranspiration in Alpine River Basins. Journal of Irrigation and Drainage Engineering 2012;138:592–9. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000453.

[26] Cobaner M, Citakoglu H, Haktanir T, Kisi O. Modifying Hargreaves-Samani equation with meteorological variables for estimation of reference evapotranspiration in Turkey. Hydrology Research 2017;48:480–97. https://doi.org/10.2166/nh.2016.217.

[27] Luo Y, Chang X, Peng S, Khan S, Wang W, Zheng Q, et al. Short-term forecasting of daily reference evapotranspiration using the Hargreaves–Samani model and temperature forecasts. Agric Water Manag 2014;136:42–51. https://doi.org/10.1016/J.AGWAT.2014.01.006.

[28] Quej VH, Almorox J, Arnaldo JA, Moratiel R. Evaluation of Temperature-Based Methods for the Estimation of Reference Evapotranspiration in the Yucatán Peninsula, Mexico. J Hydrol Eng 2018;24:05018029. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001747.

[29] Gavilán P, Lorite IJ, Tornero S, Berengena J. Regional calibration of Hargreaves equation for estimating reference ET in a semiarid environment. Agric Water Manag 2006;81:257–81. https://doi.org/10.1016/J.AGWAT.2005.05.001.

[30] Maestre-Valero JF, Martínez-Álvarez V, González-Real MM. Regionalization of the Hargreaves coefficient to estimate long-term reference evapotranspiration series in SE Spain. Spanish Journal of Agricultural Research 2013;11:1137–52. https://doi.org/10.5424/SJAR/2013114-3869.

[31] Estévez J, Gavilán P, Giráldez J V. Guidelines on validation procedures for meteorological data from automatic weather stations. J Hydrol (Amst) 2011;402:144–54. https://doi.org/10.1016/J.JHYDROL.2011.02.031.

[32] Meek DW, Hatfield JL. Data quality checking for single station meteorological databases. Agric For Meteorol 1994;69:85–109. https://doi.org/10.1016/0168-1923(94)90083-3.

[33] Zhao L, Zhao X, Zhou H, Wang X, Xing X. Prediction model for daily reference crop evapotranspiration based on hybrid algorithm and principal components analysis in Southwest China. Comput Electron Agric 2021;190:106424. https://doi.org/10.1016/J.COMPAG.2021.106424.

[34] Bellido-Jiménez JA, Estévez J, García-Marín AP. Automation of the features selection process to improve temperature-based solar radiation models in Southern Spain. 8th International Conference on Time Series and Forecasting (ITISE 2022), Gran Canaria: 2022.

[35] Fraihat H, Almbaideen AA, Al-Odienat A, Al-Naami B, De Fazio R, Visconti P. Solar Radiation Forecasting by Pearson Correlation Using LSTm Neural Network and ANFIS Method: Application in the West-Central Jordan. Future Internet 2022, Vol 14, Page 79 2022;14:79. https://doi.org/10.3390/FI14030079.

[36] Gocić M, Motamedi S, Shamshirband S, Petković D, Ch S, Hashim R, et al. Soft computing approaches for forecasting reference evapotranspiration. Comput Electron Agric 2015;113:164–73. https://doi.org/10.1016/J.COMPAG.2015.02.010.

[37] Rosenblatt F. Perceptron Simulation Experiments. Proceedings of the IRE 1960;48:301–9. https://doi.org/10.1109/JRPROC.1960.287598.

[38] Patel D, Patel S, Patel P, Shah M. Solar radiation and solar energy estimation using ANN and Fuzzy logic concept: A comprehensive and systematic study. Environmental Science and Pollution Research 2022;29:32428–42. https://doi.org/10.1007/S11356-022-19185-Z/TABLES/4.

[39] Antonopoulos VZ, Papamichail DM, Aschonitis VG, Antonopoulos A V. Solar radiation estimation methods using ANN and empirical models. Comput Electron Agric 2019;160:160–7. https://doi.org/10.1016/J.COMPAG.2019.03.022.

[40] Bellido-Jiménez JA, Estévez Gualda J, García-Marín AP. Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions. Appl Energy 2021;298. https://doi.org/10.1016/j.apenergy.2021.117211.

[41] Bellido-jiménez JA, Gualda JE, García-marín AP. Assessing machine learning models for gap filling daily rainfall series in a semiarid region of spain. Atmosphere (Basel) 2021;12. https://doi.org/10.3390/atmos12091158.

[42] Parviz L, Rasouli K. Development of Precipitation Forecast Model Based on Artificial Intelligence and Subseasonal Clustering. J Hydrol Eng 2019;24:04019053. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001862.

[43] Ghazvinian H, Bahrami H, Ghazvinian H, Heddam S. Simulation of Monthly Precipitation in Semnan City Using ANN Artificial Intelligence Model. Journal of Soft Computing in Civil Engineering 2020;4:36–46. https://doi.org/10.22115/SCCE.2020.242813.1251.

[44] Ghazvinian M, Zhang Y, Seo DJ, He M, Fernando N. A novel hybrid artificial neural network - Parametric scheme for postprocessing medium-range precipitation forecasts. Adv Water Resour 2021;151:103907. https://doi.org/10.1016/J.ADVWATRES.2021.103907.

[45] Bellido-Jiménez JA, Estévez J, García-Marín AP. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agric Water Manag 2021;245. https://doi.org/10.1016/j.agwat.2020.106558.

[46] Bellido-Jiménez JA, Estévez J, García-Marín AP. A regional machine learning method to outperform temperature-based reference evapotranspiration estimations in Southern Spain. Agric Water Manag 2022;274. https://doi.org/10.1016/j.agwat.2022.107955.

[47] Bellido-Jiménez JA, Estévez J, Vanschoren J, García-Marín AP. AgroML: An Open-Source Repository to Forecast Reference Evapotranspiration in Different Geo-Climatic Conditions Using Machine Learning and Transformer-Based Models. Agronomy 2022;12. https://doi.org/10.3390/agronomy12030656.

[48] Achite M, Jehanzaib M, Sattari MT, Toubal AK, Elshaboury N, Wałęga A, et al. Modern Techniques to Modeling Reference Evapotranspiration in a Semiarid

Area Based on ANN and GEP Models. Water 2022, Vol 14, Page 1210 2022;14:1210. https://doi.org/10.3390/W14081210.

[49] Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. J Hydrol (Amst) 2019;572:556–70. https://doi.org/10.1016/J.JHYDROL.2019.03.028.

[50] Ferreira LB, da Cunha FF. New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 2020;234:106113. https://doi.org/10.1016/J.AGWAT.2020.106113.

[51] Ferreira LB, da Cunha FF. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. Comput Electron Agric 2020;178:105728. https://doi.org/10.1016/J.COMPAG.2020.105728.

[52] Huang G Bin, Zhu QY, Siew CK. Extreme learning machine: A new learning scheme of feedforward neural networks. IEEE International Conference on Neural Networks - Conference Proceedings 2004;2:985–90. https://doi.org/10.1109/IJCNN.2004.1380068.

[53] Huang G Bin, Babri HA. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. IEEE Trans Neural Netw 1998;9:224–9. https://doi.org/10.1109/72.655045.

[54] Zhang H, Zhang S, Yin Y. Online sequential ELM algorithm with forgetting factor for real applications. Neurocomputing 2017;261:144–52. https://doi.org/10.1016/J.NEUCOM.2016.09.121.

[55] Mateo F, Carrasco JJ, Millán-Giraldo M, Sellami A, Escandell-Montero P, Martínez-Martínez JM, et al. Temperature Forecast in Buildings Using Machine Learning Techniques n.d.

[56] Dash Y, Mishra SK, Panigrahi BK. Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques. 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017 2018;2018-January:1714–8. https://doi.org/10.1109/ICICICT1.2017.8342829.

[57] Zhang X, Zhao D, Wang T, Wu X, Duan B. A novel rainfall prediction model based on CEEMDAN-PSO-ELM coupled model. Water Supply 2022;22:4531–43. https://doi.org/10.2166/WS.2022.115.

[58] Zhu B, Feng Y, Gong D, Jiang S, Zhao L, Cui N. Hybrid particle swarm optimization with extreme learning machine for daily reference evapotranspiration prediction from limited climatic data. Comput Electron Agric 2020;173:105430. https://doi.org/10.1016/J.COMPAG.2020.105430.

[59] Wu L, Zhou H, Ma X, Fan J, Zhang F. Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: Application in contrasting climates of China. J Hydrol (Amst) 2019;577:123960. https://doi.org/10.1016/J.JHYDROL.2019.123960.

[60] Fan J, Yue W, Wu L, Zhang F, Cai H, Wang X, et al. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agric For Meteorol 2018;263:225–41. https://doi.org/10.1016/J.AGRFORMET.2018.08.019.

[61] Feng Y, Cui N, Zhao L, Hu X, Gong D. Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China. J Hydrol (Amst) 2016;536:376–83. https://doi.org/10.1016/J.JHYDROL.2016.02.053.

[62] Feng Y, Hao W, Li H, Cui N, Gong D, Gao L. Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renewable and Sustainable Energy Reviews 2020;118:109393. https://doi.org/10.1016/J.RSER.2019.109393.

[63] Shamshirband S, Mohammadi K, Yee PL, Petković D, Mostafaeipour A. A comparative evaluation for identifying the suitability of extreme learning machine to predict horizontal global solar radiation. Renewable and Sustainable Energy Reviews 2015;52:1031–42. https://doi.org/10.1016/J.RSER.2015.07.173.

[64] Karaman ÖA, Tanyıldızı Ağır T, Arsel İ. Estimation of solar radiation using modern methods. Alexandria Engineering Journal 2021;60:2447–55. https://doi.org/10.1016/J.AEJ.2020.12.048.

[65] Vapnik V. The Nature of Statistical Learning Theory. vol. 8, 2000, p. 1–15. https://doi.org/10.1007/978-1-4757-3264-1_1.

[66] Fan M, Sharma A. Design and implementation of construction cost prediction model based on SVM and LSSVM in industries 4.0. International Journal of Intelligent Computing and Cybernetics 2020;14:145–57. https://doi.org/10.1108/IJICC-10-2020-0142/FULL/XML.

[67] Otchere DA, Arbi Ganat TO, Gholami R, Ridha S. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. J Pet Sci Eng 2021;200:108182. https://doi.org/10.1016/J.PETROL.2020.108182.

[68] Kurani A, Doshi P, Vakharia A, Shah M. A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting. Annals of Data Science 2023;10:183–208. https://doi.org/10.1007/S40745-021-00344-X/TABLES/2.

[69] Wang H, Hu D. Comparison of SVM and LS-SVM for regression. Proceedings of 2005 International Conference on Neural Networks and Brain Proceedings, ICNNB’05 2005;1:279–83. https://doi.org/10.1109/ICNNB.2005.1614615.

[70] Hussein E, Ghaziasgar M, Thron C. Regional rainfall prediction using support vector machine classification of large-scale precipitation maps. Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020 2020. https://doi.org/10.23919/FUSION45008.2020.9190285.

[71] Wang Y, Liu J, Li R, Suo X, Lu EH. Medium and Long-term Precipitation Prediction Using Wavelet Decomposition-prediction-reconstruction Model. Water Resources Management 2022;36:971–87. https://doi.org/10.1007/S11269-022-03063-X/TABLES/2.

[72] Fan J, Wu L, Ma X, Zhou H, Zhang F. Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renew Energy 2020;145:2034–45. https://doi.org/10.1016/J.RENENE.2019.07.104.

[73] Álvarez-Alvarado JM, Ríos-Moreno JG, Obregón-Biosca SA, Ronquillo-Lomelí G, Ventura-Ramos E, Trejo-Perea M. Hybrid Techniques to Predict Solar Radiation Using Support Vector Machine and Search Optimization Algorithms: A Review. Applied Sciences 2021, Vol 11, Page 1044 2021;11:1044. https://doi.org/10.3390/APP11031044.

[74] Guermoui M, Gairaa K, Boland J, Arrif T. A Novel Hybrid Model for Solar Radiation Forecasting Using Support Vector Machine and Bee Colony Optimization Algorithm: Review and Case Study. Journal of Solar Energy Engineering, Transactions of the ASME 2021;143. https://doi.org/10.1115/1.4047852/1085484.

[75] Ehteram M, Singh VP, Ferdowsi A, Mousavi SF, Farzin S, Karami H, et al. An improved model based on the support vector machine and cuckoo algorithm for simulating reference evapotranspiration. PLoS One 2019;14:e0217499. https://doi.org/10.1371/JOURNAL.PONE.0217499.

[76] Chia MY, Huang YF, Koo CH. Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. Comput Electron Agric 2020;175:105577. https://doi.org/10.1016/J.COMPAG.2020.105577.

[77] Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. J Hydrol (Amst) 2019;572:556–70. https://doi.org/10.1016/J.JHYDROL.2019.03.028.

[78] Weerts HJP, Mueller AC, Vanschoren J. Importance of Tuning Hyperparameters of Machine Learning Algorithms 2020.

[79] Nourani V, Elkiran G, Abdullahi J. Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. J Hydrol (Amst) 2019;577:123958. https://doi.org/10.1016/J.JHYDROL.2019.123958.

[80] Fan J, Yue W, Wu L, Zhang F, Cai H, Wang X, et al. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agric For Meteorol 2018;263:225–41. https://doi.org/10.1016/J.AGRFORMET.2018.08.019.

[81] Parker WS. Ensemble modeling, uncertainty and robust predictions. Wiley Interdiscip Rev Clim Change 2013;4:213–23. https://doi.org/10.1002/WCC.220.

[82] Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Eng Appl Artif Intell 2022;115:105151. https://doi.org/10.1016/J.ENGAPPAI.2022.105151.

[83] Nourani V, Elkiran G, Abdullahi J. Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. J Hydrol (Amst) 2019;577:123958. https://doi.org/10.1016/J.JHYDROL.2019.123958.

[84] Ferreira LB, da Cunha FF. New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 2020;234:106113. https://doi.org/10.1016/J.AGWAT.2020.106113.

[85] Liashchynskyi P, Liashchynskyi P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS 2019. https://doi.org/10.48550/arxiv.1912.06059.

[86] Bergstra J, Ca JB, Ca YB. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 2012;13:281–305.

[87] Borji A, Itti L. Bayesian optimization explains human active search. Adv Neural Inf Process Syst 2013;26.

[88] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.

[89] Hutter F, Kotthoff L, Vanschoren J. Automated Machine Learning. vol. 1. Springer International Publishing; 2019. https://doi.org/10.1007/978-3-030-05318-5.

Chapter 2

Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment

Javier Estévez [1,2,*], Juan Antonio Bellido-Jiménez[1], Xiaodong Liu[2] and Amanda Penélope García-Marín[1]

[1] Engineering Projects Area, Department of Rural Engineering, University of Córdoba, Córdoba, Spain;

[2] School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, UK;

Abstract

Accurate forecast of hydrological data such as precipitation is critical in order to provide useful information for water resources management and play a key role in different sectors. Traditional forecasting methods present many limitations due to the high-stochastic property of precipitation and its strong variability in time and space: not identifying non-linear dynamics or not solving the instability of local weather situations. In this work, several alternative models based on the combination of wavelet analysis (multiscalar decomposition) with artificial neural networks have been developed and evaluated at sixteen locations in Southern Spain (semiarid region of Andalusia), representative of different climatic and geographical conditions. Based on the capability of wavelets to describe non-linear signals, ten wavelet neural network models (WNN) have been applied to predict monthly precipitation by using short-term thermo-pluviometric time series. Overall, the forecasting results show differences between the ten models, although an effective performance (i.e., correlation coefficients ranged from 0.76 to 0.90 and Root Mean Square Error values ranged from 6.79 to 29.82 mm) was obtained at each of the locations assessed. The most appropriate input variables to obtain the best forecasts are analyzed, according to the geo-climatic characteristics of the sixteen sites studied.

1. Introduction

Precipitation, besides being one of the most important variables in hydrological models (infiltration, soil loss, droughts, overland flow production, floods, etc.), is crucial in sectors such as agriculture, tourism or even in the energy sector [1], where the absence of water can lead to the closure of nuclear plants, such as the recent case in July 2019 in France. Therefore, the improvement of precipitation predictions is one of the greatest current challenges of the scientific community. Likewise, accurate precipitation forecasting is a very difficult and relevant mechanism of the hydrologic cycle due to its high spatial-temporal variability. Because of the large number of interconnected variables that are involved in the physical modeling of precipitation, forecasting rainfall is exceptionally complicated [2]. Due to the nonlinear and dynamic characteristics of precipitation, methods like numerical weather prediction (NWP) models or even statistical models still have difficulties to provide satisfactory precipitation forecasts [3]. This is mainly due to the fact that they are subject to many uncertainties [4–9] such as not solving the local weather situations or not identifying non-linear dynamics in time series, among others [10–12].

In this sense, the mathematical models called Artificial Neural Networks (ANN), which are inspired by how the human nervous system works, have many strengths. One of them, which is highly important, is their ability to learn from experience. ANN models are based on a set of processing elements called neurons and they can accumulate a large number of behaviors, allowing users to forecast previously nonexistent patterns. Another advantage is that neurons in ANNs work in a parallel processing mechanism, being able to process—as singular or multi-layered information—big data efficiently. Lastly, they can extract complex nonlinear relationships between variables, which can be very useful for precipitation modeling. The concept of artificial neurons was introduced by the authors in [13] but the ANN applications have increased since the back-propagation learning method was developed [14]. Since then, the use of ANN in the field of research has turned into a multitude of satisfactory solutions to problems that are not easily solved with traditional techniques, especially when the quality is doubtful and the quantity is scarce [15]. One of the most used ANN architectures is the so-called feed-forward multilayer perceptron (FFMLP), where all the information propagates in one direction toward the output layer with no feedback. This architecture is explained in detail in Section 2.2. In addition, their use is very advantageous, of great versatility and easy handling because these models do not need to formulate the mathematical description of the complex mechanisms involved in the process.

In hydrological modeling, Artificial Neural Network techniques were applied for the first time by [16]. Since then, numerous works successfully address improvements in models of rainfall-runoff [17–19], stream-flow [20–22], water quality [15,23,24], groundwater [25,26], and even for data validation as a quality assurance procedure [27,28]. In 2000, the American Society of Civil Engineering published two technical works related to Hydrology and ANNs [29,30] whose results have been discussed in depth and compared to other modeling techniques. Recently, a work summarizing a review of neural network techniques applied to hydrological systems has been reported [31].

Concerning works that exclusively deal with the forecast of precipitation time series using ANN, several studies can be found in the scientific literature. An ANN model for precipitation forecasting in Thailand was developed by [32] using various meteorological parameters measured at surrounding stations. In some regions of Greece, researchers [33] obtained precipitation predictions using ANNs and 115-year datasets. Other works such as [34] and [35], used various climate indices (North Atlantic Oscillation -NAO-, Southern Oscillation -SOI-, etc.) as input variables in Korea and Australia, respectively. In China, several works based on ANNs have been developed using long-term historical datasets [3,36,37]. Moreover, similar models have been applied in different Indian regions [38–41]. Some of the main problems of this kind of model are the non-availability of historical records at many locations, the non-existence of neighboring stations, and the impossibility of arranging the previously mentioned climate indices (NAO, SOI, etc.) in near-real time in order to forecast one-step ahead.

1.1. Wavelet Multiscale Analysis

The multiscalar characterization of precipitation has been studied for several years in different regions of the world using various approaches and for different purposes [42–47]. Especially in the current context of climate variability and change, all the techniques that are capable of deepening the stochastic behavior of precipitation time series are of great interest for use in many applications [48]. One of the most effective is the wavelet analysis [49], because it can provide an exact location of any changes in the dynamic patterns of the time series, being widely applied in hydrological topics such as forecasting [50–52], rainfall trends [53] or water quality modeling [54], among others. Wavelets are a class of functions that cut up data into different frequency components and they are used to localize a given function in both position and scaling. A wavelet transformation is a powerful mathematical signal processing tool, able to produce both time and frequency information and provide multiresolution analysis. There are two main types of wavelet transforms: continuous and discrete, being the most extensively used. The main advantage of wavelets versus Fourier analysis is its power to process non-stationarity signals, determining the temporal variation of the frequency content and allowing users to track the evolution of processes at different timescales in data sequences

Different wavelet families have been studied for different purposes depending on the time series to be analyzed: Coiflets, Symlets, Daubechies, Feyer-Korovkin, and BiorSplines, among others. In hydrologic modeling, Daubechies wavelet [55] is one of the most employed due to its orthonormality properties and its good trade-off between parsimony and information plenitude [56–58]. This kind of wavelet has associated subclasses (db1 or haar, db2, db3, …, dbN) depending on the number of vanishing moments and there is a scaling function generating an orthogonal multi-resolution analysis. This multiple-level decomposition process estimates the discrete wavelet transform coefficients, breaking down the original time series into several lower-resolution components as a set of sub-signals: approximation (cAN) and details (cDN). For example, for the level of decomposition = 2 this iterative process will lead to cA2, cD1, cD2 sub-series. The approximation coefficients were produced by a low-pass filter and the details coefficients by a high-pass filter, representing the low and high-frequency components, respectively. Figure 2. 1 shows the multiresolution analysis based on this wavelet decomposition. Thus, these meteorological sub-series generated by wavelet transformation can be used as input variables in ANN approaches, giving rise to a type of so-called hybrid models: Wavelet Neural Networks (WNN).

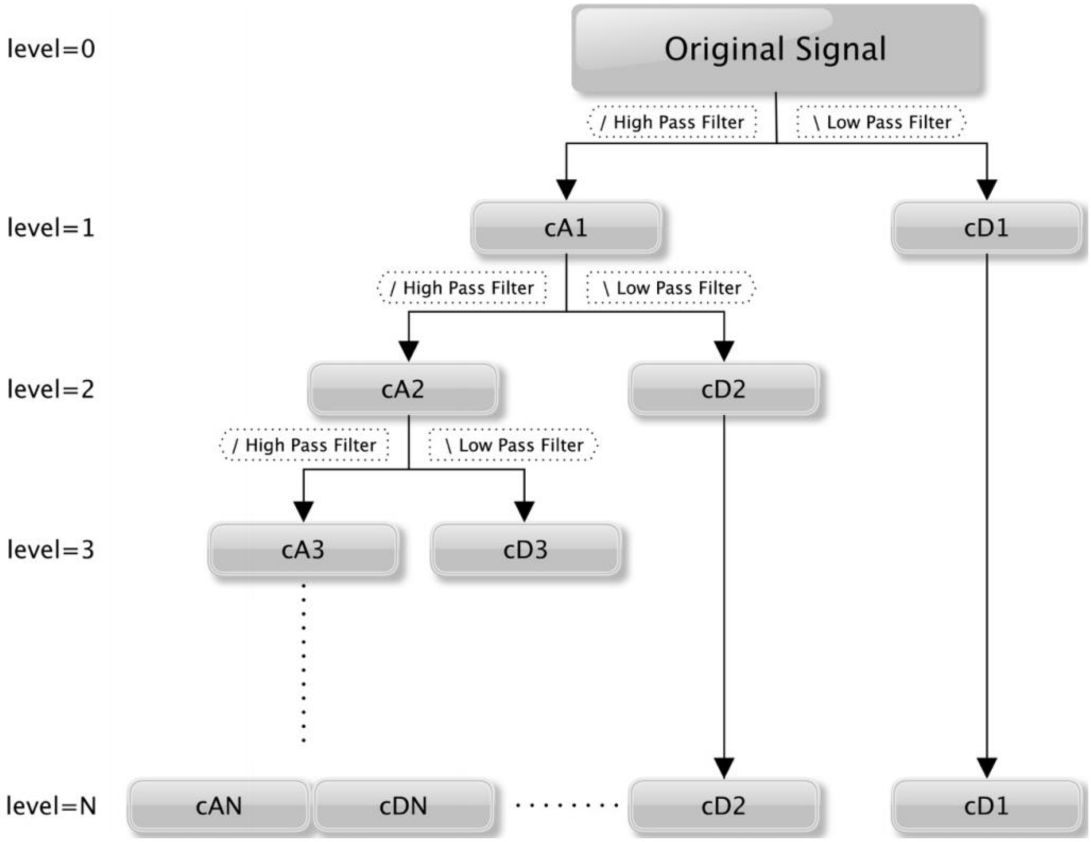


Figure 2. 1. Wavelet multiresolution analysis of original time series.

1.2.Availability of Short-Term Meteorological Series

Precipitation, and also temperature, are meteorological variables widely measured worldwide in comparison to solar radiation, humidity, or wind speed, among others [59–62]. Besides, their behavior within the climate system is being studied all over the world [63], as both variables represent the key controlling factors in the spatial variation of terrestrial ecosystem carbon exchange [64]. However, long-term series are not easily available and often contain many gaps, and have undergone homogenization or filling-gap processes usually due to changes in location, sensor replacement, variations in the mechanisms of data collection and measurements, etc

In order to improve the weather monitoring systems among other aims, the installation of automated weather stations networks able to collect at least temperature and precipitation data has been increasing since the end of the past century practically worldwide [65] and more recently with the combination of low-cost sensors and Internet of Things devices [66]. Therefore, there is currently a large availability of thermo-precipitation records from numerous spatially distributed locations with almost entirely no gaps and more than a decade in length. Thus, and due to many recent works reporting the improvement of ANN-based hydrological models combining them with wavelet analysis [3,67–70], the main goal of this work is the development and assessment of different hybrid WNN models to accurately forecast monthly precipitation in the semiarid and heterogeneous region of Andalusia (Southern Spain) using only short-term thermo-precipitation validated datasets. Due to the importance of precipitation forecasts and since the availability of these data will increase in the coming years, the present work may be extensible to many other climatic areas of the world where these records are collected. Moreover, this work evaluates the use of new input thermal variables, in addition to precipitation, to deepen the knowledge and analyze the effectiveness of these hybrid models to forecast monthly precipitation in a geo-climatic variety of locations that have very different precipitation patterns.

For these purposes, different stations in the semiarid region of Andalusia (Southern Spain) were selected. Wavelet decompositions were applied to initial datasets in order to generate the input variables of the neural network models. The performance of all the WNN approaches has been evaluated using different statistics at each location.

2. Materials and methods

2.1.Source of data

Datasets used in this work were obtained from the Agroclimatic Information Network of Andalusia and they are easily downloadable on a daily basis from http://www.juntadeandalucia.es/ agriculturaypesca/ifapa/ria/ (access on 2 August 2019), where there are some automated weather stations recently installed and others not operational. Andalusia is a semiarid region located in the South of the Iberian Peninsula (South-western Europe) covering almost 88,000 km2 and is divided into eight provinces: Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, and Sevilla. According to its relief, it is a very heterogeneous region: from the extensive coastal plains of the Guadalquivir River (at sea level) to the highest areas of the Iberian Peninsula (‘Sierra Nevada’ in the province of Granada). In terms of dryness, high contrasts are found from the Tabernas desert (province of Almería) to the rainiest areas of Spain in the ‘Sierra de Grazalema’ Natural Park (province of Cádiz). Another singularity is that it is surrounded by the Mediterranean Sea and the Atlantic Ocean at its Southeast and Southwest sides, respectively. The geographical distribution of the stations used in this work is shown in Figure 2. 2 and Table 2. 1 reports some of their characteristics, with latitudes ranging from 36.3372◦ to 38.0806◦ N, longitudes from 1.8831◦ to 7.2469◦ W and site elevations from 26 to 822 m above mean sea level. In general, the aridity increases from East (Huelva province) to West (Almería province) across the Andalusia region [71]. These sites were selected in order to represent this climatic variability of the region, including coastal (‘Málaga’ and ‘Conil de la Frontera’ stations) and inland locations, ensuring that the available time series are complete and gap-free.

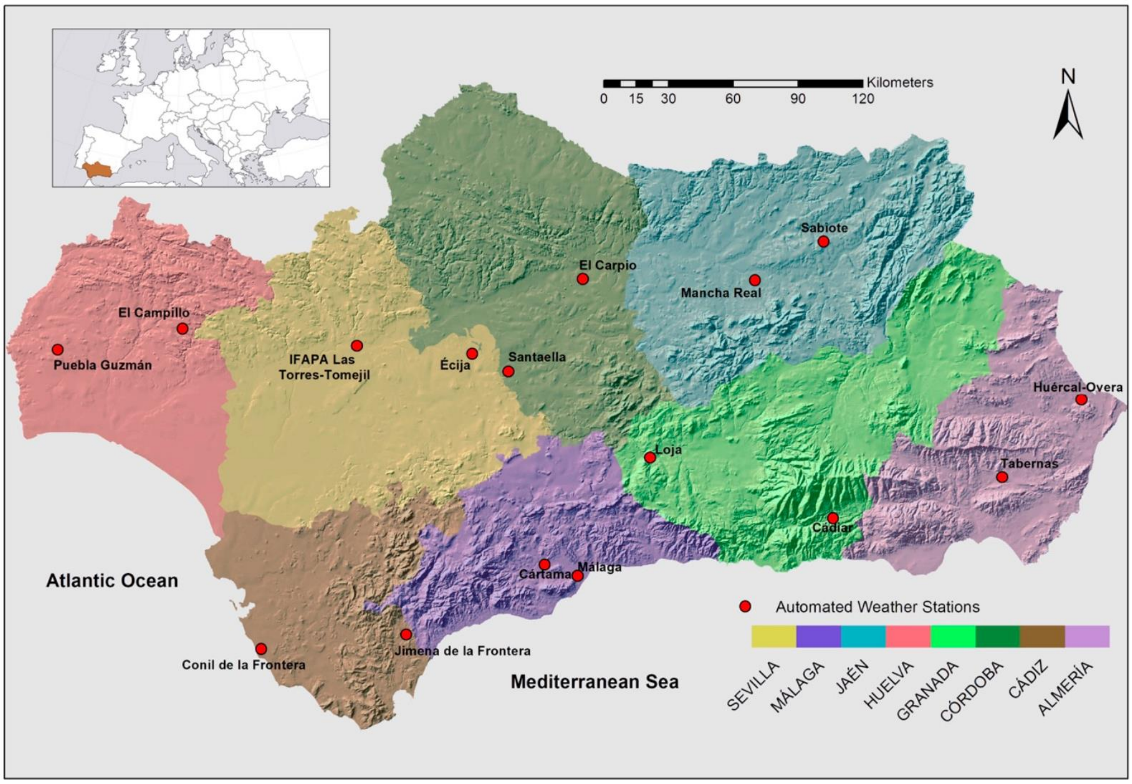


Figure 2. 2. Geographical distribution of the automated weather stations used in this work (Andalusia region - Southern Spain).

Time periods of monthly precipitation, maximum and minimum temperature datasets from each station are summarized in Table 2. 1. All of them end in July 2019 and start in 2000/2001, ranging from 213 months at 'IFAPA las Torres-Tomejil' station to 234 months at 'Huércal-Overa' station. In order to assess model performances and follow the method previously described [54], the first 85% of datasets were used to calibrate the models and the remaining 15% of the records were used for validation (at least two and a half years at all locations). Table 2. 2 shows the statistical values of these datasets for monthly precipitation, the maximum and minimum temperature for each location

In order to ensure the reliability of datasets, a set of checking quality procedures has been applied to precipitation and temperature daily data following the guidelines proposed by [72]. In addition, a specific algorithm for detecting spurious precipitation signals [73] and the spatial regression test [74] were also carried out. The application of these quality assurance techniques to hydro-meteorological data has been successfully carried out under different climatic conditions worldwide as a pre-requisite before their use [75–78]

Table 2. 1. Name of the station, province, coordinates, elevation and data time-period of the weather stations used in this study (Southern Spain)

Station Name	Province	Latitude (º)	Longitude (º)	Elevation (m)	Time Period (Calibration) Time Period (Validation)
Tabernas (ALM04)	Almería	37.0925 N	2.3011 W	435	March 2000–August 2016 September 2016–July 2019
Huércal Overa (ALM07)	Almería	37.4133 N	1.8831 W	317	February 2000–August 2016 September 2016–July 2019
Conil Frontera (CAD05)	Cádiz	36.3372 N	6.1306 W	26	November 2000–November 2016 December 2016–July 2019
Jimena Frontera (CAD07)	Cádiz	36.4136 N	5.3844 W	53	January 2001–September 2016 October 2016–July 2019
El Carpio (COR05)	Córdoba	37.9150 N	4.5025 W	165	December 2000–September 2016 November 2016–July 2019
Santaella (COR07)	Córdoba	37.5236 N	4.8842 W	207	November 2000–November 2016 December 2016–July 2019
Loja (GRA03)	Granada	37.1706 N	4.1369 W	487	October 2000–September 2016 October 2016–July 2019
Cádiar (GRA07)	Granada	36.9242 N	3.1825 W	950	October 2000–September 2016 October 2016–July 2019
Puebla Guzmán (HUE07)	Huelva	37.5533 N	7.2469 W	288	December 2000–September 2016 November 2016–July 2019
El Campillo (HUE08)	Huelva	37.6622 N	6.5981 W	406	December 2000–September 2016 November 2016–July 2019
Mancha Real (JAE04)	Jaén	37.9175 N	3.5950 W	436	October 2000–September 2016 October 2016–July 2019

Sabiote (JAE07)	Jaén	38.0806 N	3.2342 W	822	October 2000–September 2016 October 2016–July 2019
Málaga (MAG01)	Málaga	36.7575 N	4.5364 W	68	November 2000–November 2016 December 2016–July 2019
Cártama (MAG09)	Málaga	36.7181 N	4.6769 W	95	August 2001–October 2016 November 2016–July 2019
Écija (SEV07)	Sevilla	37.5942 N	5.0756 W	125	December 2000–September 2016 November 2016–July 2019
IFAPA Las Torres-Tomejil (SEV101)	Sevilla	37.4008 N	5.5875 W	75	November 2001–November 2016 December 2016–July 2019

2.2.Development of Wavelet Neural Network (WNN) Models

Several hybrid models (WNN) were developed based on the use of the sub-series resulting from the wavelet decomposition of the original series, as input variables of a feed-forward multilayer perceptron neural network (FFMLP). This architecture (Figure 2. 3) is the most widely used in water resources modeling [79] and consists of an input layer, one or more hidden layers containing network computation nodes (neurons), and the output layer that contains the target variable (predicted precipitation). The number of input nodes is equal to the number of input variables (details and approximations of sub-time series and month of year) and the number of hidden nodes is determined by trial and error procedure. One of the main keys to the good behavior of these approaches is the ability to learn from experience using the well-known backpropagation method in the training process and optimized by applying the Levenberg–Marquardt algorithm. Eventually, logarithmic sigmoidal and pure linear transfer activation functions were used for the hidden and output layers, respectively, converting input signals into output signals. Thus, the process that takes place in the neurons is the following. Firstly, the inputs are multiplied by their corresponding initial weights; these products with a bias term are summed. Afterward, this result passes as the input of an activation function which determines whether the neuron is activated or not. Then, the result advances to the next neurons and the process is repeated until the output is obtained (it is mathematically expressed as Eq. 2.1). Finally, the backpropagation training method consists of modifying the weights of the nodes based on the minimization of the bias errors (difference between target and output value) and all the process is repeated from the beginning.

Table 2. 2. Statistics of monthly precipitation, maximum and minimum temperature (Std: Standard Deviation; Max: Maximum; Min: Minimum).

Sites	Datasets	Precipitation (mm)				Maximum Temperature (°)				Minimum Temperature (°)			
		Mean	Std	Max	Min	Mean	Std	Max	Min	Mean	Std	Max	Min
Tabernas (ALM04)	All	19.95	25.56	141.40	0.00	29.85	6.59	42.55	15.53	4.69	6.40	17.18	−8.20
	Validation	18.77	27.25	141.40	0.00	29.13	6.49	41.70	17.68	4.44	6.09	15.10	−5.30
	Calibration	20.17	25.30	128.40	0.00	29.98	6.62	42.55	15.53	4.74	6.47	17.18	−8.20
Huércal-Overa (ALM07)	All	22.49	31.94	247.80	0.00	29.89	6.02	43.58	17.03	4.54	6.46	17.18	−8.85
	Validation	19.57	34.37	186.80	0.00	29.90	5.87	40.76	18.57	4.37	6.12	15.19	−6.00
	Calibration	23.02	31.55	247.80	0.00	29.88	6.06	43.58	17.03	4.58	6.53	17.18	−8.85
Conil de la Frontera (CAD05)	All	42.71	54.32	287.60	0.00	28.72	6.45	41.37	16.04	6.53	5.02	15.80	−5.38
	Validation	37.95	55.09	208.60	0.00	28.00	6.80	40.30	18.96	5.91	4.72	15.80	−1.03
	Calibration	43.58	54.28	287.60	0.00	28.86	6.39	41.37	16.04	6.65	5.07	15.37	−5.38
Jimena de la Frontera (CAD07)	All	61.05	75.03	441.00	0.00	30.18	6.74	46.57	18.64	5.99	5.26	16.02	−3.88
	Validation	63.22	86.12	371.40	0.00	29.86	5.90	42.28	19.62	5.73	5.05	14.70	−1.51
	Calibration	60.66	73.11	441.00	0.00	30.23	6.89	46.57	18.64	6.04	5.31	16.02	−3.88
El Carpio (COR05)	All	41.23	48.84	317.60	0.00	31.38	8.59	47.10	15.42	4.89	6.58	17.93	−9.54
	Validation	38.12	48.55	260.20	0.00	31.54	8.56	47.10	19.61	4.32	6.50	15.40	−6.15
	Calibration	41.78	48.99	317.60	0.00	31.35	8.61	46.94	15.42	4.99	6.60	17.93	−9.54
Santaella (COR07)	All	44.27	50.85	310.80	0.00	30.64	8.15	45.69	17.36	6.08	6.05	17.27	−8.25
	Validation	42.47	54.85	277.80	0.00	29.96	7.94	44.91	18.69	6.21	5.64	16.10	−3.05
	Calibration	44.60	50.25	310.80	0.00	30.76	8.20	45.69	17.36	6.06	6.14	17.27	−8.25
Loja (GRA03)	All	36.96	39.12	230.60	0.00	29.87	7.53	45.94	16.92	4.05	6.01	15.37	−9.45
	Validation	35.66	44.21	225.40	0.00	29.97	7.90	45.94	16.92	4.08	5.94	14.70	−5.80
	Calibration	37.20	38.25	230.60	0.00	29.86	7.48	42.85	17.08	4.05	6.04	15.37	−9.45
Cádiar (GRA07)	All	43.46	56.88	423.60	0.00	27.11	7.02	42.63	14.17	5.03	6.06	18.38	−13.30
	Validation	42.55	61.55	317.00	0.00	26.26	7.03	41.20	16.11	4.43	6.37	15.90	−13.30

	Calibration	43.62	56.18	423.60	0.00	27.26	7.03	42.63	14.17	5.14	6.02	18.38	−8.13
Puebla Guzmán (HUE07)	All	46.69	53.29	296.80	0.00	29.21	7.84	43.63	15.42	6.60	5.09	16.38	−4.02
	Validation	43.36	50.38	197.80	0.00	29.24	7.62	42.18	18.65	6.82	4.68	15.50	−0.73
	Calibration	47.29	53.90	296.80	0.00	29.21	7.89	43.63	15.42	6.56	5.17	16.38	−4.02
El Campillo (HUE08)	All	60.51	69.67	389.80	0.00	29.51	7.63	43.07	15.41	6.95	4.81	16.39	−2.39
	Validation	56.16	66.43	351.00	0.00	29.48	7.61	42.74	18.92	6.78	4.58	15.40	−1.37
	Calibration	61.28	70.38	389.80	0.00	29.51	7.65	43.07	15.41	6.98	4.86	16.39	−2.39

$$y = \emptyset\left(\sum wx + b\right) \quad \text{Eq. 2. 1}$$

where y = output value of the hidden/output node, I = input or hidden node value, ∅ = the transfer function, w = weights connecting nodes, and b = bias for each node.

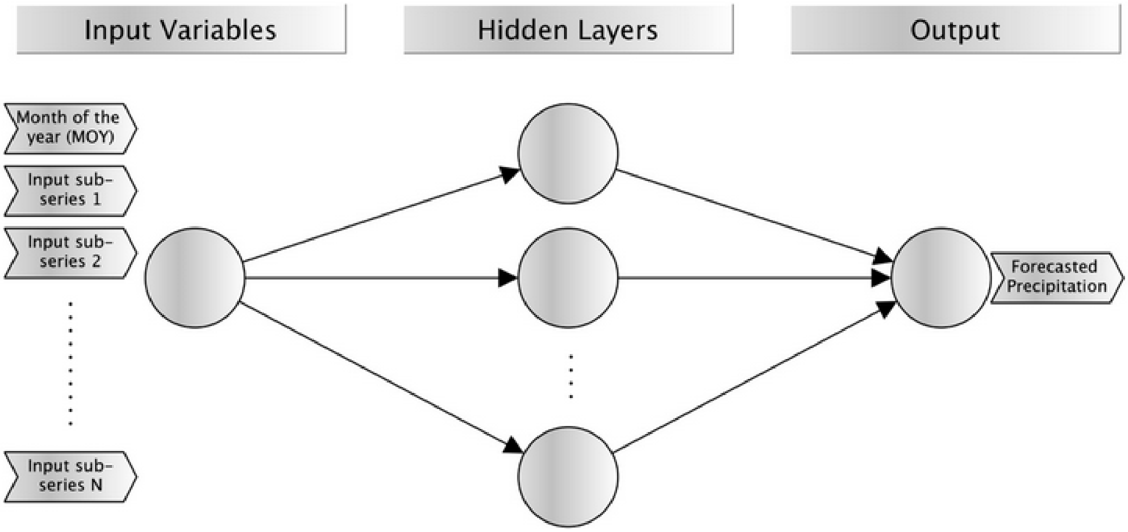


Figure 2. 3. Multilayer Perceptron Neural Network architecture used in this work

The selection of the Daubechies wavelet of order 5 (db5) was performed after a trial and error procedure checking Daubechies wavelet from order 1 to 10 [68,80,81], although similar results were found with db9. The wavelet decomposition process was carried out according to the procedure in [82] at level 3, based on the size of validation datasets for testing the model performances [69]. Finally, the optimal number of neurons in the hidden layer [2,68,83] was set to eight, after testing from two to ten in steps of one and checking the FFMLP performance.

Thus, each dataset was decomposed by wavelet transformation into sub-series containing approximation coefficients (cA3) and details coefficients (cD1, cD2, and cD3). They were used as input variables for the WNN models as well as the month of the year (MOY: 1 = January, 2 = February…12 = December), and monthly precipitation original series was used as the target output values. An example of the sub-series of precipitation (details and approximations) after the wavelet decomposition as well as the original signal is represented in Figure 2. 4 for Málaga station.

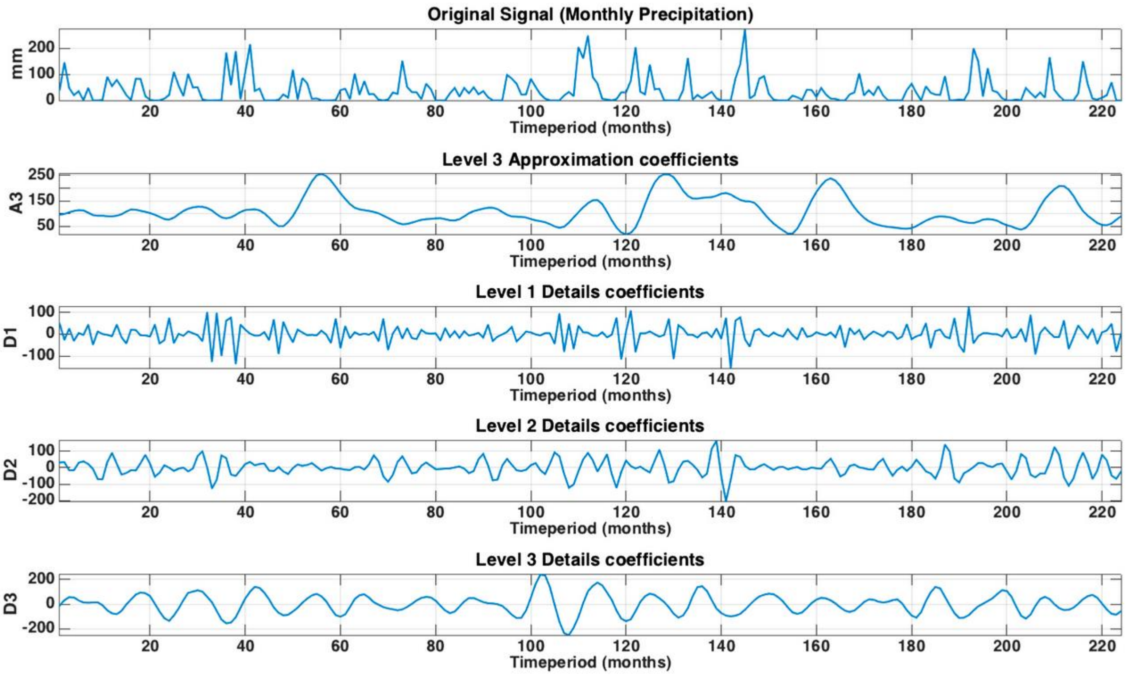


Figure 2. 4. Original values and decomposed sub-series of monthly precipitation by wavelet transformation at Málaga station (MAG01) (2001–2019)

The input variables used in each model are summarized in Table 2. 3. Inputs and the number of variables of each of the wavelet neural network models (WNN) models evaluated in this work (i = month; MOY = month of the year; P = precipitation; DTRM = mean diurnal temperature range; DTRX = maximum diurnal temperature range; DTRN = minimum diurnal temperature range; MTR=monthly temperature range; TX=maximum temperature; Tn = minimum temperature). All the models used Month of the year (MOY) and precipitation signal decomposed by wavelets transformation. The proposed models used a different combination of variables. For instance, the input variables of the Model I were MOY and monthly precipitation signal (decomposed into D1, D2, D3, and A3 coefficients). In contrast, the Model IX used MOY, precipitation signal (decomposed into D1, D2, D3, and A3 coefficients), and monthly minimum temperature signal (decomposed into D1, D2, D3, and A3 coefficients).

Table 2. 3. Inputs and the number of variables of each of the wavelet neural network models (WNN) models evaluated in this work (i = month; MOY = month of the year; P = precipitation; DTR_M = mean diurnal temperature range; DTR_X = maximum diurnal temperature range; DTR_N = minimum diurnal temperature range; MTR=monthly temperature range; T_X=maximum temperature; Tn = minimum temperature).

Models	Output	Input Variables	Nº Variables
I	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m)	5
II	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), $P_{\{decomposed\}}$ (i−1)	9
III	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), $DTR_{M\ \{decomposed\}}$ (i)	9
IV	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), $DTR_{X\ \{decomposed\}}$ (i)	9
V	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), $DTR_{N\ \{decomposed\}}$ (i)	9

VI	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), $DTR_{X\ \{decomposed\}}$ (mi $DTR_{N\ \{decomposed\}}$ (mi	13
VII	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), MTR $_{\{decomposed\}}$ (i)	9
VIII	P (i + 1)	MOY, $P_{\{decomposed\}}$ (m), $T_{x\{decomposed\}}$ (i)	9
IX	P (m + 1)	MOY, $P_{\{decomposed\}}$ (m), $T_{n\{decomposed\}}$ (m)	9
X	P (m + 1)	MOY, $P_{\{decomposed\}}$ (m), $T_{x\{decomposed\}}$, $T_{n\{decomposed\}}$ (m)	13

2.3. Statistical Analysis and Performance Criteria

In order to evaluate the performance of different models developed in this work, forecasted and measured precipitation values were compared by using simple error analysis. Thus, common statistical indices widely used to assess hydro-meteorological prediction models [26,61,68] were estimated: RMSE (root mean square error), R (Correlation Coefficient), MAPE (mean absolute percentage error) and NSE (Nash–Sutcliffe model efficiency coefficient, also known as the coefficient of efficiency). These statistics are summarized from Equations 2.2 to 2.5:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2} \qquad \text{Eq. 2. 2}$$

$$R = \frac{\sum_{i=1}^{m}(meas_i - \overline{meas})(pred_i - \overline{pred})}{\sqrt{\sum_{i=1}^{m}(meas_i - \overline{meas})^2 \sum_{i=1}^{n}(pred_i - \overline{pred})^2}} \qquad \text{Eq. 2. 3}$$

$$MAPE = \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{meas_i - pred_i}{meas_i}\right| \qquad \text{Eq. 2. 4}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(meas_i - pred_i)^2}{\sum_{i=1}^{n}(meas_i - \overline{meas})^2} \qquad \text{Eq. 2. 5}$$

where the m is the number of months and $meas_i$, $pred_i$, $\overline{meas}$, $\overline{pred}$ are the precipitation measured at month i, precipitation forecasted at month I, the mean of measured monthly precipitation and the mean of forecasted monthly precipitation, respectively

In addition, two performance measures were also carried out: Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). These indices have the singularity of considering the number of trained parameters and they are based on the parsimony. AIC and BIC were initially reported by [84] and [85], respectively, and they have been frequently used for assessing hydrological models [86–88]. Both expressions are described in Equations 2.6 and 2.7:

$$AIC = m\ln(RMSE) + 2p \qquad \text{Eq. 2. 6}$$

$$BIC = m\ln(RMSE) + p\ln(m) \qquad \text{Eq. 2. 7}$$

where p is the number of free parameters in each model (the total amount of weights and biases), being the best model performance the one with lowest AIC and BIC values. These indices deal with the trade-off between the prediction error (RMSE) and the complexity of the model, combining a term reflecting how well the forecasts fit the data with a term penalizing the model in proportion to its number of estimated parameters [89].

3. Results and Discussion

3.1. Pre-Processing Input Datasets

Validated daily records (precipitation, maximum and minimum temperature) obtained after the application of quality control procedures were used to create different monthly datasets. Monthly precipitation (P) values were used as an input in all the models assessed. Apart from max/min monthly temperature records (T_x and T_n, respectively), various temperature-based monthly time series were also created from daily values: mean daily temperature range (DTRm), maximum daily temperature range (DTR_X), minimum daily temperature range (DTRn) and monthly temperature range (MTR). Daily temperature range (DTR) is the difference between daily maximum temperature and daily minimum temperature, with DTRm, DTRx, DTRn, being the mean, maximum, and minimum DTR measured in a month, respectively. MTR is obtained as the difference between the maximum and minimum temperature measured on a monthly basis.

3.2. Performance of the Models

In general, regarding forecasted validation datasets and the common statistics, Model X was one of the best performers in most of the locations studied, although Model I showed the best results, on average, of BIC and AIC indices (Figure 2. 5), followed by Models II, IX, VIII, V, IV, III, VII, X, and VI. The minimum values obtained for both indices by using Models I, II, IX, and X were registered in the driest location (Tabernas station), in Conil de la Frontera by using Model III and Model VII, in Mancha Real by using Model IV, in IFAPA-Las Torres station by using Model VI and in Huércal-Overa station by using Model V and Model VIII. As in the results reported by [87], both indices produced the same model selection, with the exception of Model VII which showed the best AIC and BIC performances in Sabiote and Conil de la Frontera stations, respectively. Overall, the results from BIC and AIC values indicated a worse performance of the approaches that use more variables (Model VI and Model X) than the rest, with Model I being the one with the lowest indices. Thus, the number of estimated parameters (weights and biases) in each of the models played a determining role in these indices.

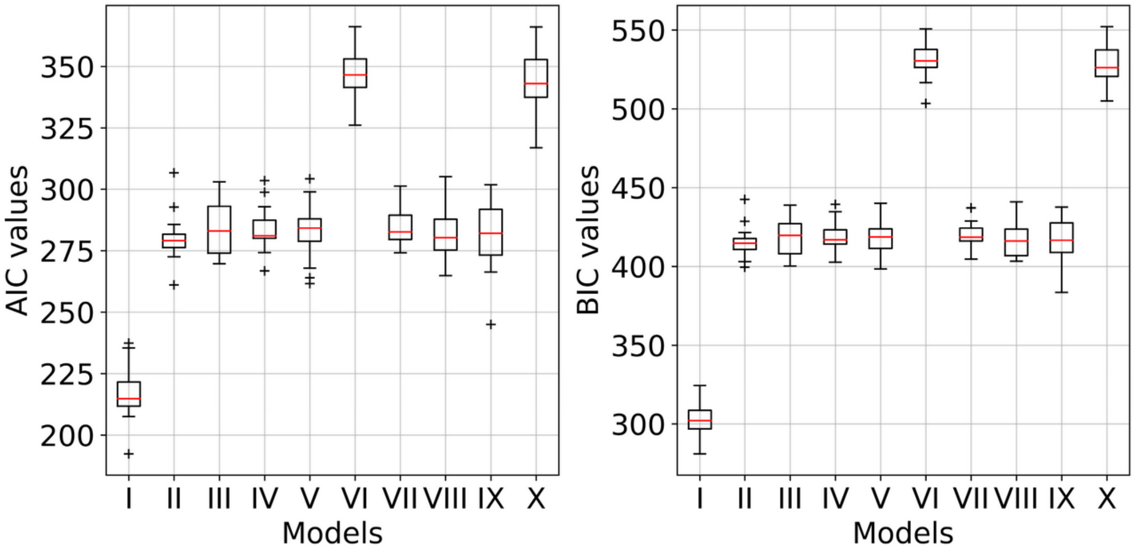


Figure 2. 5. Box-plot of the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values obtained by using the ten models (validation datasets) for all the sites studied. On each box: the red central mark=median; bottom and top edges of the box

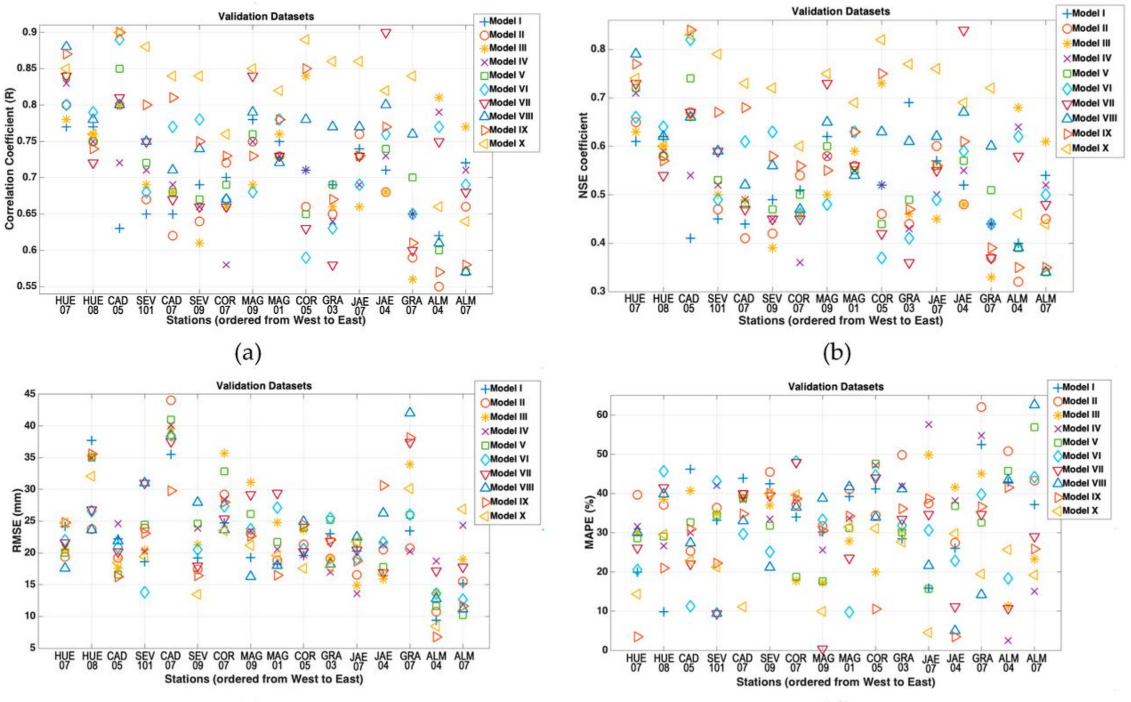
In terms of the statistics R, RMSE, MAPE, and NSE, the mean, maximum, and minimum values obtained in the sixteen locations are summarized in Table 2. 4 for each model and dataset studied. Regarding validation forecasts, Model I obtained the best R (0.78) and NSE (0.62) values in Cártama station, and the lowest RMSE and MAPE values in Tabernas (9.39 mm) and El Campillo (9.82%) stations, respectively. On average, Model I had a generally better performance than other related models carried out in Greece [33] or in Jordan [83], but with R and NSE values lower than those reported by [68] in one station in India. However, Model II was the one that showed the worst results in almost all sites and for all the statistics studied, although with some exceptions. These results indicated that for the goal of this work, the information contained in the 'two months before' precipitation signal is not as relevant as the one contained in the 'one month before' signal. Model III had, on average, a slightly better performance, registering the lowest MAPE and RMSE values in Tabernas station (11.39% and 13.75 mm, respectively) and the best R (0.84) and NSE (0.73) values in El Carpio station. However, Model IV obtained good statistical indices in Cádiar, Mancha Real, and Almería stations, while Model V gave the lowest RMSE (10.20 mm) in the Huércal-Overa station. In general, the mean results obtained by using the variables DTR_m (Model III), DTR_x (Model IV), and DTR_n (Model V) were similar and better than those reported by [33] and [83], although in terms of MAPE, Model IV gave the best values in the most arid sites (Tabernas and Huércal-Overa stations). The next model assessed (VI) had a good performance in the two coastal locations: Conil de la Frontera station (highest R = 0.89 and NSE = 0.82 values) and Málaga station (best MAPE value = 9.80%), which may indicate that the joint use of DTR_x and DTR_n variables in areas near the sea could be recommended. Model VII gave the best MAPE values of all the models and sites in Cártama (0.40%) and IFAPA-Las Torres Tomejil (9.44%) stations and the best R (0.90), RMSE (16.95 mm) and NSE (0.84) values in Sabiote station, indicating that the new input variable MTR can be very useful in some locations. Finally, Model VIII (using T_x)

obtained the lowest RMSE value in Huércal-Overa (11.16 mm), the best MAPE value in Sabiote (4.96%), and the highest R (0.88), and NSE (0.79) values in El Campillo station, where Model IX (using T_n) also obtained the lowest MAPE value (3.45%). In addition, this model (IX) had a very good behavior also in Sabiote (MAPE = 3.51%), Conil de la Frontera (R = 0.90 and NSE = 0.84), El Carpio (R = 0.85 and NSE = 0.75) and Tabernas (RMSE = 6.79 mm) stations. Regarding these last two models, no clear improvement was observed to recommend Model VII or Model IX based on the geo-climatic conditions. On average, the highest values of R (0.82) and NSE (0.69) were obtained by Model X (using T_x and T_n) for the validation dataset and for all the sites, ranging from R = 0.90 and NSE = 0.83 (Conil de la Frontera station) to R = 0.64 and NSE = 0.44 (Húercal Overa station). In general, using Model II the lowest average values of R (0.69) and NSE (0.50) were given, and also the minimum values obtained for all the sites (R = 0.55 and NSE = 0.32 in Tabernas station). Regarding RMSE average values, they ranged from 21.49 (Model X) to 31.55 mm (Model II), while the highest value (44.03 mm) was registered in Jimena de la Frontera station by using also Model II, with this station being the one with the rainiest month (371.40 mm). Attending to MAPE average values, Model X was able to forecast with the lowest error (23.61%) followed by Model VII (28.02 %), ranging from 4.57% (Mancha Real station) to 40.04% (Écija station) and from 0.40% (Cártama station) to 47.94% (Santaella station), respectively. Instead, Model II gave the highest MAPE average value (39.93%) as well as the greatest percentage registered from all the stations (62.02%) in Cádiar (the highest location). As in other related works [3,32,34,68], a better general performance in calibration datasets can be observed.

Table 2. 4. Summary of correlation coefficient (R), root mean square error (RMSE), mean absolute percentage error (MAPE) and Nash–Sutcliffe model efficiency coefficient (NSE) values for all the models assessed.

Models	**Datasets**	**R Max/Mean/Min**	**RMSE(mm) Max/Mean/Min**	**MAPE(%) Max/Mean/Min**	**NSE Max/Mean/Min**
I	Validation	0.78/0.70/0.62	9.39/21.69/37.74	9.82/33.94/52.52	0.62/0.51/0.40
	Calibration	0.83/0.74/0.65	11.75/20.67/29.60	9.86/16.07/22.57	0.81/0.72/0.63
II	Validation	0.80/0.69/0.55	10.73/31.55/44.03	25.34/39.93/62.02	0.67/0.50/0.32
	Calibration	0.98/0.92/0.79	11.89/16.18/29.21	1.86/7.84/22.99	0.96/0.85/0.63
III	Validation	0.84/0.71/0.56	13.75/24.17/39.53	11.39/31.57/49.86	0.73/0.54/0.33
	Calibration	0.95/0.92/0.87	11.33/17.59/26.97	4.92/8.63/15.91	0.91/0.84/0.75
IV	Validation	0.83/0.71/0.58	13.61/23.25/40.12	2.50/34.84/57.58	0.71/0.52/0.36
	Calibration	0.92/0.85/0.74	11.12/16.84/24.50	4.11/8.21/17.00	0.91/0.85/0.73
V	Validation	0.85/0.71/0.57	10.20/23.68/41.00	15.73/33.04/56.89	0.74/0.53/0.34
	Calibration	0.97/0.93/0.85	11.54/15.66/24.80	1.58/6.50/16.68	0.94/0.87/0.73
VI	Validation	0.89/0.73/0.59	12.64/22.48/38.51	9.80/31.19/48.17	0.82/0.55/0.37
	Calibration	0.97/0.95/0.91	7.79/13.96/18.28	0.12/5.05/11.89	0.95/0.90/0.82
VII	Validation	0.90/0.72/0.58	16.95/24.44/37.55	0.40/28.02/47.94	0.84/0.55/0.36
	Calibration	0.97/0.95/0.92	8.48/14.65/23.19	1.67/4.46/9.58	0.95/0.90/0.85
VIII	Validation	0.88/0.75/0.57	11.16/22.86/42.04	4.96/32.37/62.61	0.79/0.58/0.34
	Calibration	0.98/0.94/0.91	7.67/15.34/25.52	0.02/4.23/9.05	0.96/0.89/0.83
IX	Validation	0.90/0.74/0.57	6.79/22.84/38.17	3.45/28.05/41.50	0.84/0.58/0.35
	Calibration	0.97/0.94/0.88	8.02/15.03/21.22	1.67/5.09/11.15	0.94/0.89/0.77
X	Validation	0.90/0.82/0.64	8.49/21.49/38.39	4.57/23.61/40.04	0.83/0.69/0.44
	Calibration	0.98/0.94/0.90	9.61/14.61/20.88	2.45/5.71/11.40	0.96/0.89/0.81

In order to evaluate the results obtained by using the ten models at each location, the statistical indices R, NSE, RMSE, and MAPE are shown in Figure 2. 6 (a, b, c, and d, respectively) for validation datasets. In Figure 2. 6a, it can be observed that in the most humid site (Puebla-Guzmán station = HUE07), located in the western region of Andalusia, the highest R (0.88) and NSE (0.79) values were obtained by Model VIII, followed by IX, X, and VII. The other station situated in Huelva province (El Campillo station = HUE08) registered very homogeneous values of R and NSE by using all the models, with Model VI being the best one with values of 0.79 and 0.64, respectively. One of the best correlation coefficients and NSE values were obtained in Conil de la Frontera (CAD05) by using Model VI (R = 0.89 and NSE = 0.82), Model IX (R = 0.90 and NSE = 0.84) and Model X (R = 0.90 and NSE = 0.83). In this coastal location, Models IV and I gave the worst values. However, Model X was the best one for the following stations: IFAPA-Las Torres (SEV101), Jimena de la Frontera (CAD07), Écija (SEV09), Santaella (COR07), Cártama (MAG09), Málaga (MAG01), El Carpio (COR05), Loja (GRA03), Mancha Real (JAE07) and Cádiar (GRA07) stations (from West to East). Finally, for the driest locations (ALM04 = Tabernas and ALM07 = Huércal-Overa), situated in the eastern part of Andalusia, the model that derived the best R and NSE indices was the Model III, the one using DTR_m as an input variable. Therefore, these results indicate that the use of the DTR_m signal could be recommended for precipitation forecasting in arid stations. Considering Figure 2. 6c, for the stations located in Huelva province (western part of Andalusia), the lowest RMSE values were obtained by Model VIII in HUE07 (17.60 mm) and HUE08 (23.62 mm), which could indicate the suitability of using this model in the less arid areas of Southern Spain. The location with the highest RMSE value was the rainiest site: Jimena de la Frontera (CAD07), while the lowest ones (6.79 and 10.20 mm) were obtained at the most arid stations by using Models IX (ALM04) and V (ALM07), respectively. Finally, MAPE values (Figure 2. 6d) showed high variability between stations and also for the different models evaluated. The highest range between the best and the worst models was obtained in Mancha Real (JAE07), while the most homogeneous values occurred in Loja (GRA03). On average, the worst MAPE values were obtained in the highest location (Cádiar = GRA07), but no relationship was found between elevation and MAPE. For all the locations studied, several models were able to obtain MAPE values lower than 25%, including excellent performances such as those given by Model IX in Puebla Guzmán (HUE07), Model VII in Cártama (MAG09) or Model IV in Tabernas (ALM04), with the exception of Model X in Loja station (GRA03) obtaining 27.61%.



(a) (b)

(c) (d)

Figure 2. 6. Results of the statistical performance obtained at each of the 16 locations studied: (a) R; (b) NSE; (c) RMSE; (d) MAPE.

Finally, measured and forecasted values of monthly precipitation at four stations (Conil de la Frontera, Tabernas, Loja, and Sabiote) during calibration and validation periods are represented in Figure 2. 7. When attending to the validation datasets, a very good performance of Model VI can be observed in a coastal location such as Conil de la Frontera (CAD05), using MOY, precipitation, DTRx, and DTRn as input variables and obtaining R = 0.89 and MAPE = 11.29%. In addition, this model also gave the lowest percentages of error at Málaga (MAG01) coastal station (MAPE = 9.80%). Thus, the input variables used in this model were more efficient at coastal locations than other variable combinations in terms of predictability performance. Slightly worse was the behavior of Model III (MOY, precipitation, and DTR_m as input variables) in Tabernas (the driest station), with R=0.81, and MAPE = 11.39%, but being able to properly forecast the peak of 141.40 mm. Likewise, the validation period results obtained in Loja station (GRA03) by applying Model X indicated, in general, a satisfactory performance in terms of R (0.86), RMSE (17.81 mm), and NSE (0.72), although the peak of 225.40 mm was not predicted so accurately. Finally, the modeled datasets using Model VII in Sabiote station (JAE04) are represented. Regarding the validation period, the values of NSE, R, and RMSE obtained with this model showed the best model performance in this site (0.84, 0.90, and 16.95 mm, respectively) and also gave an acceptable MAPE of 11.18%. Furthermore, this model that used MOY, precipitation, and MTR as input variables, forecasted with lowest MAPE values in another two interior stations: Cártama (MAG09) and IFAPA-Las Torres Tomejil (SEV101), although its performance was not so good in other inland locations.

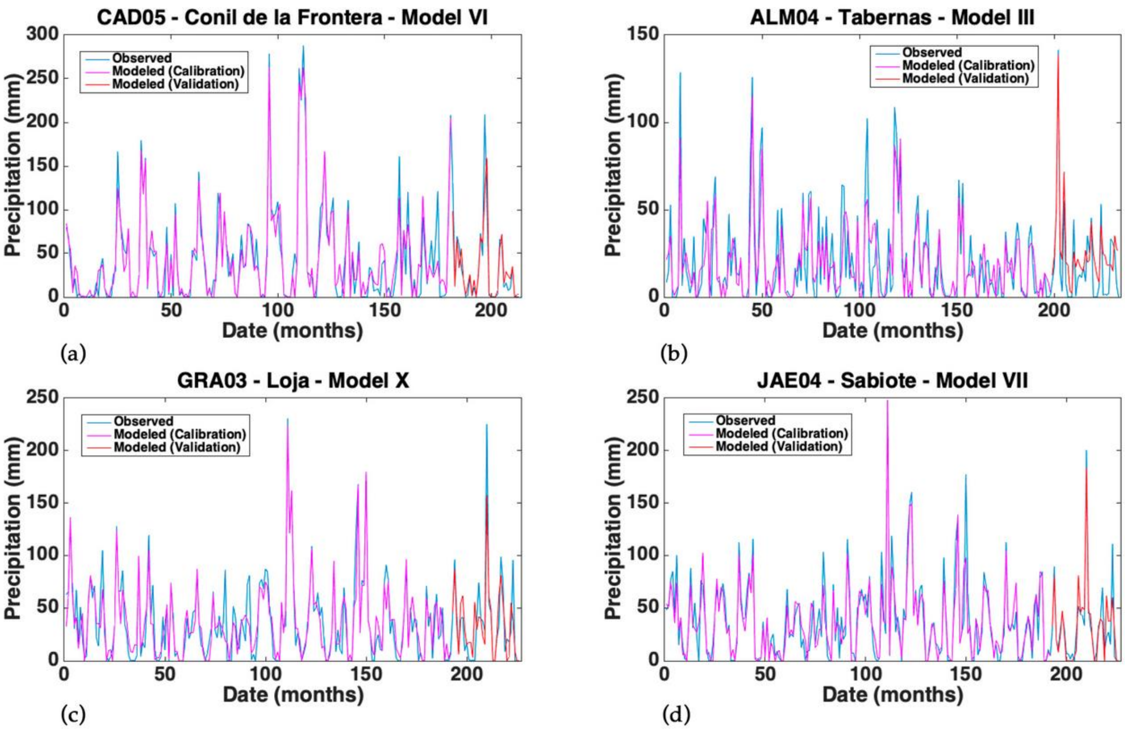


Figure 2. 7. Plot of measured and forecasted monthly precipitation at four stations: Conil de la Frontera (a), Tabernas (b), Loja (c), and Sabiote (d) using Models VI, III, X and VIII, respectively.

From these results, it has been verified that the introduction of easily estimated input variables such as DTR_x, DTR_n, DTR_m, MTR or MOY into WNN models is very useful for improving precipitation predictions one month ahead, especially when there is no availability of long-term datasets. In general, the results obtained by applying the proposed models in all stations in Southern Spain provided better RMSE values than the best of several WNN monthly precipitation models assessed by [68] at one station located in the east of India and also better than those reported by [3] at 24 locations in China, with both works needing the use of long-term historical series. Moreover, RMSE values were also lower in this work than the reported by [2] in ten stations in Guilin (China) using evolutionary models. In terms of efficiency, mean NSE values indicated a good degree of efficiency for all the models, being much higher than the values reported by [90] in Iran using ANN to predict monthly precipitation using 30-year series. RMSE values obtained with ANN models by [90] were worse than those given with the ten approaches assessed in this work. In addition, the correlation coefficients obtained in this work in all locations except at Huércal-Overa and Santaella sites were better than those reported by [33] in four stations in Greece for cumulative four-month precipitation predictions using ANN models. Regarding this statistic, the best result reported by [83] for the monthly precipitation in one of the three stations studied in Jordan was similar to the best values obtained in Santaella and Huércal-Overa stations but lower than those given in the rest of the locations. However, the correlation coefficient obtained by [90] with ANN and singular spectrum analysis model was better than the average performance of all the models, although models from V to X gave higher R values at least in one location of the sixteen sites evaluated.

4. Conclusions

Different configurations of a hybrid model combining wavelet analysis and artificial neural network for time series forecasting of monthly precipitation have been developed and assessed at sixteen locations in Southern Spain (semiarid region). The main novelty of this work is the use of thermal variables, besides precipitation, never used before, such as the daily and monthly thermal range, as well as the month of the year, the use of short-term time series, and the application to datasets from sixteen sites having very different climatic and geographical conditions. Firstly, a set of sub-signals were obtained from original validated datasets carrying out a multilevel decomposition process by wavelet transformation. Then, these new time series and months of the year were used as input variables of the ten models evaluated, with the original monthly precipitation being the output variable. The models were calibrated using the first 85% of the datasets and the rest of the data were used for model validation (at least two and a half years at all locations). The results indicated that nonlinear dynamics of the different thermal variables used and also precipitation were properly characterized by wavelet decomposition in order to satisfactorily forecast precipitation one month ahead, although the performance of the models was not the same for the different locations evaluated. For each location, it was found that there were at least one or more models with acceptable statistical performance ($R > 0.76$; $NSE > 0.60$; $RMSE < 29.82$ mm and $MAPE < 27.62\%$).

In general, the model that used precipitation, maximum and minimum temperature (X) had the best statistical performance in most of the locations studied. However, the model using precipitation and the mean diurnal temperature range (III) gave the best results at the most arid sites. Regarding coastal locations, the lowest mean absolute percentage of errors was obtained by the model using precipitation, maximum, and minimum diurnal temperature range (VI). By contrast, the model using only precipitation signal (I) obtained the best BIC at all locations and the lowest AIC values at twelve sites due to the reduced number of input variables but did not get the best results in any other statistical indices except in El Campillo station, the second rainiest site of this study. Although no relationship between model performance and site elevation was found, the worst mean absolute percentage error was obtained in the highest site studied (Cádiar station). Finally, the model using precipitation and monthly temperature range (VII) gave satisfactory results in terms of predictability error in three interior locations. Therefore, an overall analysis of the general results obtained in this work indicates the suitability of the type of input variables used in WNN models that accurately describe precipitation processes according to geo-climatic characteristics.

Since most of the thermo-pluviometric sensors installed on automatic weather stations networks worldwide do not have long-term time series and considering that precipitation is a meteorological variable with high spatial variability, these types of models are of great interest to the monthly precipitation forecast in locations where only short length records are available. Further works using different artificial intelligence approaches such as support vector machines or extreme learning machines may be carried out to compare the performance of these kinds of models once they are joined to wavelet analysis.

5. References

[1] Linnerud K, Mideksa T, Journal GE-TE, 2011 undefined. The impact of climate change on nuclear power supply. IaeeOrg n.d.

[2] Jiang L, Wu J. Hybrid PSO and GA for neural network evolutionary in monthly rainfall forecasting. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2013;7802 LNAI:79–88. https://doi.org/10.1007/978-3-642-36546-1_9/COVER.

[3] Liu Q, Zou Y, Liu X, Linge N. A survey on rainfall forecasting using artificial neural network. Int J Embed Syst 2019;11:240–9. https://doi.org/10.1504/IJES.2019.098300.

[4] Jabbari A, Bae DH. Application of Artificial Neural Networks for Accuracy Enhancements of Real-Time Flood Forecasting in the Imjin Basin. Water 2018, Vol 10, Page 1626 2018;10:1626. https://doi.org/10.3390/W10111626.

[5] Alotaibi K, Ghumman AR, Haider H, Ghazaw YM, Shafiquzzaman M. Future Predictions of Rainfall and Temperature Using GCM and ANN for Arid Regions: A Case Study for the Qassim Region, Saudi Arabia. Water 2018, Vol 10, Page 1260 2018;10:1260. https://doi.org/10.3390/W10091260.

[6] Moghim S, Bras RL. Bias Correction of Climate Modeled Temperature and Precipitation Using Artificial Neural Networks. J Hydrometeorol 2017;18:1867–84. https://doi.org/10.1175/JHM-D-16-0247.1.

[7] Yang Z, Hsu K, Sorooshian S, Xu X, Braithwaite D, Verbist KMJ. Bias adjustment of satellite-based precipitation estimation using gauge observations: A case study in Chile. Journal of Geophysical Research: Atmospheres 2016;121:3790–806. https://doi.org/10.1002/2015JD024540.

[8] Crochemore L, Ramos MH, Pappenberger F. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol Earth Syst Sci 2016;20:3601–18. https://doi.org/10.5194/HESS-20-3601-2016.

[9] Valverde Ramírez MC, De Campos Velho HF, Ferreira NJ. Artificial neural network technique for rainfall forecasting applied to the São Paulo region. J Hydrol (Amst) 2005;301:146–62. https://doi.org/10.1016/J.JHYDROL.2004.06.028.

[10] Darji MP, Dabhi VK, Prajapati HB. Rainfall forecasting using neural network: A survey. Conference Proceeding - 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015 2015:706–13. https://doi.org/10.1109/ICACEA.2015.7164782.

[11] Nanda SK, Tripathy DP, Nayak SK, Mohapatra S. Intelligent Systems and Applications. Intelligent Systems and Applications 2013;12:1–22. https://doi.org/10.5815/ijisa.2013.12.01.

[12] Geetha G, Samuel Selvaraj R. Prediction of monthly rainfall in Chennai using back... - Google Scholar. International Journal of Engineering Science and Technology 3.1, 2011.

[13] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33. https://doi.org/10.1007/BF02478259/METRICS.

[14] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 323:6088 1986;323:533–6. https://doi.org/10.1038/323533a0.

[15] Maier HR, Dandy GC. The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters. Water Resour Res 1996;32:1013–22. https://doi.org/10.1029/96WR03529.

[16] French MN, Krajewski WF, Cuykendall RR. Rainfall forecasting in space and time using a neural network. J Hydrol (Amst) 1992;137:1–31. https://doi.org/10.1016/0022-1694(92)90046-X.

[17] Senthil Kumar AR, Sudheer KP, Jain SK, Agarwal PK. Rainfall-runoff modelling using artificial neural networks: comparison of network types. Hydrol Process 2005;19:1277–91. https://doi.org/10.1002/HYP.5581.

[18] Fernando DAK, Jayawardena AW. Runoff Forecasting Using RBF Networks with OLS Algorithm. J Hydrol Eng 1998;3:203–9. https://doi.org/10.1061/(ASCE)1084-0699(1998)3:3(203).

[19] Dawson CW, Wilby R. An artificial neural network approach to rainfall-runoff modelling. Https://DoiOrg/101080/02626669809492102 2009;43:47–66. https://doi.org/10.1080/02626669809492102.

[20] Jeong D Il, Kim YO. Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. Hydrol Process 2005;19:3819–35. https://doi.org/10.1002/HYP.5983.

[21] Riad S, Mania J, Bouchaou L, Najjar Y. Predicting catchment flow in a semi-arid region via an artificial neural network technique. Hydrol Process 2004;18:2387–93. https://doi.org/10.1002/HYP.1469.

[22] Birikundavyi S, Labib R, Trung HT, Rousselle J. Performance of Neural Networks in Daily Streamflow Forecasting. J Hydrol Eng 2002;7:392–8. https://doi.org/10.1061/(ASCE)1084-0699(2002)7:5(392).

[23] Kim RJ, Loucks DP, Stedinger JR. Artificial Neural Network Models of Watershed Nutrient Loading. Water Resources Management 2012;26:2781–97. https://doi.org/10.1007/S11269-012-0045-X/FIGURES/7.

[24] Zaheer I, Bai C-G. APPLICATION OF ARTIFICIAL NEURAL NETWORK FOR WATER QUALITY MANAGEMENT. Lowland Technology International 2003;5:10–5.

[25] Nourani V, Mousavi S. Spatiotemporal groundwater level modeling using hybrid artificial intelligence-meshless method. J Hydrol (Amst) 2016;536:10–25. https://doi.org/10.1016/J.JHYDROL.2016.02.030.

[26] Talei A, Chua LHC, Wong TSW. Evaluation of rainfall and discharge inputs used by Adaptive Network-based Fuzzy Inference Systems (ANFIS) in rainfall–runoff modeling. J Hydrol (Amst) 2010;391:248–62. https://doi.org/10.1016/J.JHYDROL.2010.07.023.

[27] López-Lineros M, Estévez J, Giráldez J V., Madueño A. A new quality control procedure based on non-linear autoregressive neural network for validating raw river stage data. J Hydrol (Amst) 2014;510:103–9. https://doi.org/10.1016/J.JHYDROL.2013.12.026.

[28] Sciuto G, Bonaccorso B, Cancelliere A, Rossi G. Quality control of daily rainfall data with neural networks. J Hydrol (Amst) 2009;364:13–22. https://doi.org/10.1016/J.JHYDROL.2008.10.008.

[29] Hydrology ATC on A of ANN in. Artificial Neural Networks in Hydrology. II: Hydrologic Applications. J Hydrol Eng 2000;5:124–37. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(124).

[30] Hydrology ATC on A of ANN in. Artificial Neural Networks in Hydrology. I: Preliminary Concepts. J Hydrol Eng 2000;5:115–23. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(115).

[31] Oyebode O, Stretch D. Neural network modeling of hydrological systems: A review of implementation techniques. Nat Resour Model 2019;32:e12189. https://doi.org/10.1111/NRM.12189.

[32] Hung NQ, Babel MS, Weesakul S, Tripathi NK. An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrol Earth Syst Sci 2009;13:1413–25.

[33] Moustris KP, Larissi IK, Nastos PT, Paliatsos AG. Precipitation Forecast Using Artificial Neural Networks in Specific Regions of Greece. Water Resources Management 2011;25:1979–93. https://doi.org/10.1007/S11269-011-9790-5/METRICS.

[34] Lee J, Kim CG, Lee JE, Kim NW, Kim H. Application of Artificial Neural Networks to Rainfall Forecasting in the Geum River Basin, Korea. Water 2018, Vol 10, Page 1448 2018;10:1448. https://doi.org/10.3390/W10101448.

[35] Abbot J, Marohasy -. Forecasting of medium-term rainfall using Artificial Neural Networks: Case studies from Eastern Australia. Engineering and mathematical topics in rainfall, London: IntechOpen; 2018.

[36] Yang Y, Luo Y. Using the Back Propagation Neural Network Approach to Bias Correct TMPA Data in the Arid Region of Northwest China. J Hydrometeorol 2014;15:459–73. https://doi.org/10.1175/JHM-D-13-041.1.

[37] Wu X, Hongxing C, Flitman A, Fengying W, Guolin F. Forecasting Monsoon Precipitation Using Artificial Neural Networks. Adv Atmos Sci 2001;18:949–58. https://doi.org/10.1007/BF03403515/METRICS.

[38] Tyagi N, Kumar A. Comparative analysis of backpropagation and RBF neural network on monthly rainfall prediction. Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016 2016;1. https://doi.org/10.1109/INVENTIVE.2016.7823234.

[39] Manek AH, Singh PK. Comparative study of neural network architectures for rainfall prediction. Proceedings - 2016 IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development, TIAR 2016 2016:171–4. https://doi.org/10.1109/TIAR.2016.7801233.

[40] Goyal MK. Monthly rainfall prediction using wavelet regression and neural network: an analysis of 1901–2002 data, Assam, India. Theor Appl Climatol 2014;118:25–34. https://doi.org/10.1007/S00704-013-1029-3/FIGURES/4.

[41] Acharya N, Shrivastava NA, Panigrahi BK, Mohanty UC. Development of an artificial neural network based multi-model ensemble to estimate the northeast monsoon rainfall over south peninsular India: An application of extreme learning machine. Clim Dyn 2014;43:1303–10. https://doi.org/10.1007/S00382-013-1942-2/FIGURES/6.

[42] García-Marín AP, Estévez J, Morbidelli R, Saltalippi C, Ayuso-Muñoz JL, Flammini A. Assessing Inhomogeneities in Extreme Annual Rainfall Data Series by Multifractal Approach. Water 2020, Vol 12, Page 1030 2020;12:1030. https://doi.org/10.3390/W12041030.

[43] Bohlinger P, Sorteberg A, Liu C, Rasmussen R, Sodemann H, Ogawa F. Multiscale characteristics of an extreme precipitation event over Nepal. Quarterly Journal of the Royal Meteorological Society 2019;145:179–96. https://doi.org/10.1002/QJ.3418.

[44] Medina-Cobo MT, García-Marín AP, Estévez J, Jiménez-Hornero FJ, Ayuso-Muñoz JL. Obtaining Homogeneous Regions by Determining the Generalized Fractal Dimensions of Validated Daily Rainfall Data Sets. Water Resources Management 2017;31:2333–48. https://doi.org/10.1007/S11269-017-1653-2/FIGURES/4.

[45] Medina-Cobo MT, García-Marín AP, Estévez J, Ayuso-Muñoz JL. The identification of an appropriate Minimum Inter-event Time (MIT) based on multifractal characterization of rainfall data series. Hydrol Process 2016;30:3507–17. https://doi.org/10.1002/HYP.10875.

[46] García-Marín AP, Estévez J, Medina-Cobo MT, Ayuso-Muñoz JL. Delimiting homogeneous regions using the multifractal properties of validated rainfall data series. J Hydrol (Amst) 2015;529:106–19. https://doi.org/10.1016/J.JHYDROL.2015.07.021.

[47] Samuel JM, Sivapalan M. A comparative modeling analysis of multiscale temporal variability of rainfall in Australia. Water Resour Res 2008;44:7401. https://doi.org/10.1029/2007WR006373.

[48] Estévez J, García Marín A, Báez Benitez J, Casas Castillo MC, Telesca L. Introduction to the special issue on “hydro-meteorological time series analysis and their relation to climate change.” Acta Geophysica 2018;66:317–8. https://doi.org/10.1007/S11600-018-0144-Z/METRICS.

[49] Grossmann A, Morlet J. Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape. Https://DoiOrg/101137/0515056 2006;15:723–36. https://doi.org/10.1137/0515056.

[50] Sang YF. A review on the applications of wavelet transform in hydrology time series analysis. Atmos Res 2013;122:8–15. https://doi.org/10.1016/J.ATMOSRES.2012.11.003.

[51] Maheswaran R, Khosa R. Comparative study of different wavelets for hydrologic forecasting. Comput Geosci 2012;46:284–95. https://doi.org/10.1016/J.CAGEO.2011.12.015.

[52] Adamowski J, Chan HF. A wavelet neural network conjunction model for groundwater level forecasting. J Hydrol (Amst) 2011;407:28–40. https://doi.org/10.1016/J.JHYDROL.2011.06.013.

[53] Baddoo TD, Guan Y, Zhang D, Andam-Akorful SA. Rainfall Variability in the Huangfuchuang Watershed and Its Relationship with ENSO. Water 2015, Vol 7, Pages 3243-3262 2015;7:3243–62. https://doi.org/10.3390/W7073243.

[54] Wang Y, Yuan Y, Pan Y, Fan Z. Modeling Daily and Monthly Water Quality Indicators in a Canal Using a Hybrid Wavelet-Based Support Vector Regression Structure. Water 2020, Vol 12, Page 1476 2020;12:1476. https://doi.org/10.3390/W12051476.

[55] Lindley D V. Ten lectures on wavelets. SIAM: Philadelphia, PA, USA 1992.

[56] Guimarães Santos CA, Silva GBL da. Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models.

Https://DoiOrg/101080/026266672013800944 2014;59:312–24. https://doi.org/10.1080/02626667.2013.800944.

[57] Nalley D, Adamowski J, Khalil B. Using discrete wavelet transforms to analyze trends in streamflow and precipitation in Quebec and Ontario (1954–2008). J Hydrol (Amst) 2012;475:204–28. https://doi.org/10.1016/J.JHYDROL.2012.09.049.

[58] Benaouda D, Murtagh F, Starck JL, Renaud O. Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting. Neurocomputing 2006;70:139–54. https://doi.org/10.1016/J.NEUCOM.2006.04.005.

[59] WMO. Guide to instruments and methods of observations. Geneva, Switzerland, 2018; Volume 8 2008;I & II:I.8-1 to I.9-1.

[60] De Paola F, Giugni M. Coupled Spatial Distribution of Rainfall and Temperature in USA. Procedia Environ Sci 2013;19:178–87. https://doi.org/10.1016/J.PROENV.2013.06.020.

[61] Estévez J, Padilla FLM, Gavilán P. Evaluation and Regional Calibration of Solar Radiation Prediction Models in Southern Spain. Journal of Irrigation and Drainage Engineering 2012;138:868–79. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000493.

[62] Library WO, Eccel E. Estimating air humidity from temperature and precipitation measures for modelling applications. Meteorological Applications 2012;19:118–28. https://doi.org/10.1002/MET.258.

[63] IPCC. Climate Change 2013: The Physical Science Basis. An overview of the Working Group 1 contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). EGUGA 2014;16:3544.

[64] Chen Z, Yu G, Ge J, Sun X, Hirano T, Saigusa N, et al. Temperature and precipitation control of the spatial variation of terrestrial ecosystem carbon exchange in the Asian region. Agric For Meteorol 2013;182–183:266–76. https://doi.org/10.1016/J.AGRFORMET.2013.04.026.

[65] Lewis E, Fowler H, Alexander L, Dunn R, Mcclean F, Barbero R, et al. GSDR: A Global Sub-Daily Rainfall Dataset. J Clim 2019;32:4715–29. https://doi.org/10.1175/JCLI-D-18-0143.1.

[66] Strigaro D, Cannata M, Antonovic M. Boosting a Weather Monitoring System in Low Income Economies Using Open and Non-Conventional Systems: Data Quality Analysis. Sensors 2019, Vol 19, Page 1185 2019;19:1185. https://doi.org/10.3390/S19051185.

[67] Wei S, Yang H, Song J, Abbaspour K, Xu Z. A wavelet-neural network hybrid modelling approach for estimating and predicting river monthly flows. Https://DoiOrg/101080/026266672012754102 2013;58:374–89. https://doi.org/10.1080/02626667.2012.754102.

[68] Venkata Ramana R, Krishna B, Kumar SR, Pandey NG. Monthly Rainfall Prediction Using Wavelet Neural Network Analysis. Water Resources Management 2013;27:3697–711. https://doi.org/10.1007/S11269-013-0374-4/FIGURES/9.

[69] Wu CL, Chau KW, Li YS. Methods to improve neural network performance in daily flows prediction. J Hydrol (Amst) 2009;372:80–93. https://doi.org/10.1016/J.JHYDROL.2009.03.038.

[70] Nourani V, Alami MT, Aminfar MH. A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Eng Appl Artif Intell 2009;22:466–72. https://doi.org/10.1016/J.ENGAPPAI.2008.09.003.

[71] Gómez-Zotano J, Alcántara-Manzanares J, Olmedo-Cobo JA, Martínez-Ibarra E. La sistematización del clima mediterráneo: identificación, clasificación y caracterización climática de Andalucía (España). Revista de Geografía Norte Grande 2015;2015:161–80. https://doi.org/10.4067/S0718-34022015000200009.

[72] Estévez J, Gavilán P, Giráldez J V. Guidelines on validation procedures for meteorological data from automatic weather stations. J Hydrol (Amst) 2011;402:144–54. https://doi.org/10.1016/J.JHYDROL.2011.02.031.

[73] Estévez J, Gavilán P, García-Marín AP, Zardi D. Detection of spurious precipitation signals from automatic weather stations in irrigated areas. International Journal of Climatology 2015;35:1556–68. https://doi.org/10.1002/JOC.4076.

[74] Estévez J, Gavilán P, García-Marín AP. Spatial regression test for ensuring temperature data quality in southern Spain. Theor Appl Climatol 2018;131:309–18. https://doi.org/10.1007/S00704-016-1982-8/TABLES/2.

[75] Nourani V, Elkiran G, Abdullahi J. Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. J Hydrol (Amst) 2019;577:123958. https://doi.org/10.1016/J.JHYDROL.2019.123958.

[76] Islam AT, Shen S, Yang S, Hu Z, Chu R. Assessing recent impacts of climate change on design water requirement of Boro rice season in Bangladesh. Theor Appl Climatol 2019;138:97–113. https://doi.org/10.1007/S00704-019-02818-8/TABLES/4.

[77] Yi Z, Zhao H, Jiang Y. Continuous Daily Evapotranspiration Estimation at the Field-Scale over Heterogeneous Agricultural Areas by Fusing ASTER and MODIS Data. Remote Sensing 2018, Vol 10, Page 1694 2018;10:1694. https://doi.org/10.3390/RS10111694.

[78] Estévez J, García-Marín AP, Morábito JA, Cavagnaro M. Quality assurance procedures for validating meteorological input variables of reference evapotranspiration in mendoza province (Argentina). Agric Water Manag 2016;172:96–109. https://doi.org/10.1016/J.AGWAT.2016.04.019.

[79] Wang W, Gelder PHAJMV, Vrijling JK, Ma J. Forecasting daily streamflow using hybrid ANN models. J Hydrol (Amst) 2006;324:383–99. https://doi.org/10.1016/J.JHYDROL.2005.09.032.

[80] Pal L, Chandra |, Ojha SP, Surendra |, Chandniha K, Kumar | Amit, et al. Regional scale analysis of trends in rainfall using nonparametric methods and wavelet transforms over a semi-arid region in India. International Journal of Climatology 2019;39:2737–64. https://doi.org/10.1002/JOC.5985.

[81] Shoaib M, Shamseldin AY, Melville BW. Comparative study of different wavelet based neural network models for rainfall–runoff modeling. J Hydrol (Amst) 2014;515:47–58. https://doi.org/10.1016/J.JHYDROL.2014.04.055.

[82] Du K, Zhao Y, Lei J. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. J Hydrol (Amst) 2017;552:44–51. https://doi.org/10.1016/J.JHYDROL.2017.06.019.

[83] Aksoy H, Dahamsheh A. Artificial neural network models for forecasting monthly precipitation in Jordan. Stochastic Environmental Research and Risk Assessment 2009;23:917–31. https://doi.org/10.1007/S00477-008-0267-X/TABLES/6.

[84] Akaike H. A New Look at the Statistical Model Identification. IEEE Trans Automat Contr 1974;19:716–23. https://doi.org/10.1109/TAC.1974.1100705.

[85] Rissanen J. Modeling by shortest data description. Automatica 1978;14:465–71. https://doi.org/10.1016/0005-1098(78)90005-5.

[86] Nourani V, Komasi M. A geomorphology-based ANFIS model for multi-station modeling of rainfall–runoff process. J Hydrol (Amst) 2013;490:41–55. https://doi.org/10.1016/J.JHYDROL.2013.03.024.

[87] Laio F, Di Baldassarre G, Montanari A. Model selection techniques for the frequency analysis of hydrological extremes. Water Resour Res 2009;45:7416. https://doi.org/10.1029/2007WR006666.

[88] Dawson CW, Wilby RL. Hydrological modelling using artificial neural networks 2001.

[89] Kriegeskorte N. Crossvalidation in Brain Imaging Analysis. BioRxiv 2015:017418. https://doi.org/10.1101/017418.

[90] Kalteh AM. Enhanced Monthly Precipitation Forecasting Using Artificial Neural Network and Singular Spectrum Analysis Conjunction Models. INAE Letters 2017 2:3 2017;2:73–81. https://doi.org/10.1007/S41403-017-0025-9.

Chapter 3

New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain

Juan Antonio Bellido-Jiménez[1], Javier Estévez[1], Amanda P. García-Marín[1]

[1] Projects Engineering Area, Department of Rural Engineering, University of Córdoba, Spain

Abstract:

The estimation of Reference Evapotranspiration (ET_0) is crucial to estimate crop water requirements, especially in developing countries and areas with scarce water resources. In these regions, the impossibility of collecting all the required data to compute FAO56 Penman-Monteith equation (FAO PM) makes scientists search for new methodologies to accurately estimate ET_0 with the minimum number of climatic parameters. In this work, several neural network approaches have been evaluated for estimating ET_0 using datasets from five weather stations located in Southern Spain (the semiarid region of Andalusia). The assessment of statistical performance (Root Mean Square Error -RMSE-, Mean Bias Error -MBE-, coefficient of determination -R^2- and Nash-Sutcliffe model efficiency coefficient -NSE-) of models namely Multilayer perceptron (MLP), Generalized Regression Neural Network (GRNN), Extreme Learning Machine (ELM), Support Vector Machines (SVM), Random Forest (RF) and XGBoost were carried out using different input variables configurations. Only temperature-based data were used as inputs; the calculation of new variables called $Energy_T$ (the integral of the half-hourly temperature values of a day) and Hourmin (the difference in hours between time sunset and the time when the maximum temperature occurs) had promising results for the most humid stations. The good results obtained with $Energy_T$ when it is used as an input of the system demonstrated that the information contained in it gives a detailed characterization of the daily thermic behavior at each location, resulting in a more efficient model than those using only daily maximum, minimum temperature, and extraterrestrial radiation values. In general, the modeling results showed that no model firmly outperformed the others, although MLP and ELM were commonly the models that gave the best performances for all sites: mean values of R^2 >0.89, mean values of NSE

>0.88, mean values of RMSE <0.67 mm/day and mean values of MBE ranging from -0.17 to 0.30 mm/day. Therefore, $Energy_T$ and Hourmin can be used to estimate ET_0 more accurately in stations where data acquisition is limited, like in developing countries or at low-cost weather stations that cannot collect all the required meteorological variables used in FAO56 PM. Overall, the use of ELM is recommended due to its high performance in terms of efficiency (NSE) for all configurations and all locations, especially using $Energy_T$ as an input variable.

1. Introduction

Food supply guarantee for an exponential growth population under climate change is one of the major challenges to our current society. Besides, this fact is accentuated under pandemic conditions as we have suffered worldwide during a great part of 2020 due to COVID-19. For the hydrological cycle and agronomic process, Reference Evapotranspiration (ET_0) is one of the most important. Accurate calculation of ET_0 is crucial to water resource management and irrigation scheduling, especially in semiarid areas such as the Andalusia region (Southern Spain). It determines the evaporative demand of the atmosphere in a hypothetical grass reference crop with specific characteristics, which makes this parameter affected only by climatic conditions [1].

Several statistical methodologies have been studied over the world, although the FAO-56 Penman-Monteith (FAO PM) equation is considered by The Food and Agricultural Organization of the United Nations (FAO) as the sole reference mathematical method [2]. It is proven to be a high accuracy globally method in different climatic conditions [3–7]. Another well-accepted method in both practical and research applications is the ASCE-PM [8], being identical to FAO56-PM on a daily basis.

These methods are physics-based, and the required parameters are the air temperature, the relative humidity of the air, the solar radiation, the wind speed, the atmospheric pressure, and the soil heat flux. Unfortunately, the implementation and maintenance of weather stations that collect these meteorological data are very expensive, even for developed countries. Therefore, it is not always possible to find locations where all these parameters are recorded, especially true for developing countries, where reliable data on wind speed and solar radiation are missing [9–12]. Due to these restrictions, simplified statistical equations with fewer input climatic parameters can be used to estimate ET_0. Allen et al. [1] proposed that when not enough data are available to solve FAO PM equation, it could be considered using the Hargreaves-Samani (HS) equation [13], which only requires maximum and minimum daily temperature (T_x and T_n) and extraterrestrial solar radiation (R_a) (that can be approximated using latitude and the day of the year). Several articles have studied the performance of the HS model in different countries and the improvement of its accuracy with local calibration [14–20]. For example, Martínez-Cob and Tejero-Juste [16] analyzed this method for semi-arid climates in the north of Spain; Vanderlinden et al. [20] compared different approaches between inland and coastal weather stations in the south of Spain; Trajkovic [19] studied seven weather stations located in Serbia; Gavilán et al. [14] carried out a regional calibration of HS method in Andalusia region; Todorovic et al. [21] analyzed the efficiency of HS in 577 weather stations in 16 Mediterranean countries. One of the compelling reasons for using this equation extensively is that there are so many weather stations that record temperature data, and their installation and maintenance are the cheapest.

Even though HS method is the most common worldwide, there are several methodologies to calculate ET_0 using limited climatic data: Blaney [22] and Thornthwaite and Mather [23], among other equations, use temperature as the only input Brisdata to obtain ET_0. Even FAO-56 [2] contemplates a temperature-based approach using approximations to the actual vapor pressure (e_a), the solar radiation (R_s), and a local long-term average of the wind speed (u_2). Schendel [24] made a contribution using temperature and relative humidity, whereas Priestley and Taylor [25,26] and Bristow and Campbell [26] introduced solar radiation and temperature in their solutions. Numerous articles studied and compared these methods at different locations and climatic conditions as Germany [27], Ghana [28], Switzerland [29], North America Great Plains [30], and Spain [31].

In the last decade, advances in computation have led to the introduction of Machine Learning (ML) methodologies in reference to evapotranspiration calculation, proving its high accuracy results by using different approaches. Several ML models have been studied such as Multilayer Perceptron (MLP) [32–35], Support Vector Machine (SVM) [35–41], Decision Tree (DT), and ensemble learning models [35,42–45], Extreme Learning Machine (ELM) [35,46–48], Generalized Regression Neural Network (GRNN) [42,47,49], Convolutional Neural Networks (CNN[50], and another technique such as Gene Expression Programming (GEP) [51,52] and Adaptive Neuro-Fuzzy Inference System (ANFIS) [40,51–54].

The use of limited climatic data is a common practice among scientists when analyzing different ML architectures and models. In this sense, Tabari et al. [40] studied and compared daily ET_0 with ANFIS, SVMs, and Regression Analysis versus empirical models (Hargreaves-Samani, Ritchie, Jensen-Haise, Turc, Blaney-Criddle) in a semi-arid highland environment of Iran. These different models were tested with temperature-based parameters. However, to improve each model's accuracy, new climatological parameters such as solar radiation, relative humidity, and wind speed were introduced. The best results were obtained with ANFIS, and the use of mean temperature, solar radiation, relative humidity, and wind speed. Wen et al [41] used SVM and MLP to model ET_0 with limited climatic data in extremely arid regions in China (one of the most arid zones in the world), comparing results to empirical models such as Priestley-Taylor and Hargreaves. The most accurate model was SVM using temperature, solar radiation, and wind speed. Feng et al [47] analyzed the performance of the ELM and GRNN models in 6 meteorological stations of the Sichuan basin, southwest China, using only temperature data; ELM outperformed GRNN. Adamala [55] studied the performance of Wavelet Neural Networks (WNN) in different climatological regions of India using only maximum and minimum temperature and contrasted it with MLP, Linear regression(LR), and Wavelet Regression (WR); even though WNN yielded good results, MLP gave a better performance. Ferreira and da Cunha [50] compared ANN, RF, XGBoost, and CNN using different approaches of hourly temperature, relative humidity, and extraterrestrial solar radiation values, being the CNN model the most precise solution.

In general, the introduction of new climatic parameters such as relative humidity, solar radiation, and/or wind speed is a common practice to improve the accuracy in all different models and regions. These introductions are constraint to the installation of new sensors, which is very expensive in both installation and maintenance. For example, the installation of 10 new Automated Weather Stations (AWS) in Ghana in 2019 cost around 450 000 € [56]. Moreover, the technological improvements in The Internet of Things (IoT), the huge number of devices connected to the internet, and its availability with new technologies like 5G connections led to a new important role of low-cost and non-conventional devices in AWS. The data storage is cheaper nowadays, enabling AWS to increase the sampling rate from days to minutes, and the use of new available low-cost air temperature sensors with open-source hardware like Raspberry or Arduino is escalating worldwide with great results in terms of accuracy [57,58]. However, despite the existence of new sensors for other climatic parameters such as wind speed or relative humidity, their performances are not as good yet [57–59]. Therefore, due to the promising results obtained by Ferreira et al. [50] in Brazil with the use of sub-daily temperature data, the present work studies the performance of different ML models with new approaches and using novel daily parameters ($Energy_T$ and Hourmin) exclusively obtained from air half-hourly temperature measurements.

The main objectives of this work are I) the assessment of different temperature-based ML models (MLP, SVM, RF, XGBoost, GRNN, and ELM) to estimate ET_0 in different sites located in the heterogeneous climatic region of Andalusia (Southern Spain), II) the comparative analysis of the different configurations (combinations of several input variables) for each location, and III) the performance evaluation in terms of seasonality for each model. For these purposes, five locations were selected based on their different geo-climatic characteristics in the Andalusia region. Eleven configurations for each neural network model were evaluated in order to analyze their performance at each location. Finally, the temporal assessment (seasonality) of the models' performance was also studied.

It is important to note that no similar works have been carried out using this kind of method in Southern Spain. Moreover, the Bayesian optimization to determine the fittest machine learning hyperparameters/architecture has been applied as a novel and efficient method instead of the traditional trial/error procedure widely used in ET_0 modeling (i.e.: [36,47,50]).

2. Materials and methods

2.1.Source of data

This study is carried out in the region of Andalusia, located in the south-west of Europe and the south of the Iberian peninsula, ranging the meridians from 1º to 7ºW, the parallels from 37º to 39ºN and an elevation from 26 to 822 meters above mean sea level in a total area of 87.268 Km^2 (Figure 3. 1). The climate is semiarid, where inland locations have torrid summers and harsh winters, whereas, on the coast, the temperature is slightly more temperate. On the other hand, mountainous locations are rainy and quite cold during the winter [60]. The datasets used in this work belong to the Agroclimatic Information Network of Andalusia (RIAA) which can be downloaded at https://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController. Five stations were selected to represent the different variability of climatic conditions of the region in terms of the UNEP aridity index [61], ranging from 0.178 in Tabernas to 0.555 in Aroche stations, including coastal and inland locations. The coordinates and other characteristics of the automated weather stations used in this work are reported in Table 3. 1, while Figure 3. 2 shows the monthly variation of mean daily temperature, relative humidity, wind speed, and solar radiation of these weather stations.

Figure 3. 1. Spatial distribution of the five weather stations used in this work (Aroche, Conil de la Frontera, Córdoba, Málaga, and Tabernas).

Table 3. 1. Summary of weather stations sites used in this paper (Lat: Latitude; Lon: Longitude; Altitude: elevation above mean sea level, mean annual precipitation, and Aridity Index as the relation Precipitation/ET$_0$ as the annual mean of the whole dataset period)

Name	**Lat (ºN)**	**Lon (ºW)**	**Altitude (m)**	**Mean annual precipitation (mm)**	**Aridity Index**	**Time period (avail. days)**
Córdoba	37.8569	4.8027	94.0	589.52	0.4616	2000 – 2018 (6397)
Málaga	36.7564	4.5375	55.0	434.39	0.3666	2000 – 2018 (6438)
Tabernas	37.0911	2.3022	502.0	237.80	0.1780	2000 – 2018 (6694)

Conil de la Frontera	36.3327	6.1325	22.0	470.74	0.4790	2000 – 2018 (5868)
Aroche	37.9580	6.9450	293.0	632.26	0.5550	2000 – 2018 (6399)

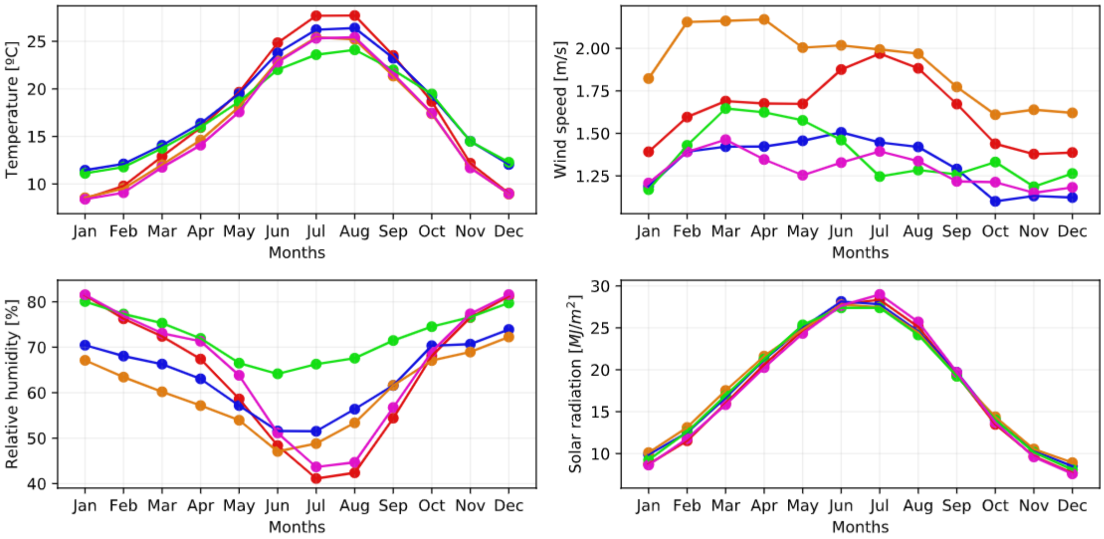


Figure 3. 2. Monthly values of temperature, relative humidity, wind speed, and solar radiation for all the weather stations (COR – Córdoba, MAG – Málaga, TAB – Tabernas, CON – Conil and ARO – Aroche).

Available datasets are divided into two main groups according to their sampling time, semi-hourly and daily sampling data. On semi-hourly datasets, also called intraday datasets, the temperature, relative humidity, wind speed, and radiation are recorded every 30 minutes. On daily data, we find the maximum, mean, and minimum temperature (Tx, Tm, Tn, respectively), the maximum, mean, and minimum relative humidity (RHx, RHm, RHn respectively), the maximum and mean wind speed (u_x, u respectively), the total solar radiation (R_s), the total precipitation (Preci) and the time of day when maximum temperature and minimum temperature occurs ($Hourmin_{Tx}$ and $Hourmin_{Tn}$). Table 3. 2 shows the statistics values of these datasets.

Table 3. 2. Statistics of maximum and minimum temperature, and ET_0 (Max: Maximum, Min: Minimum, Std: Standard Deviation.

		Maximum temperature (ºC)				**Minimum temperature (ºC)**				**ET_0 (mm/day)**			
Sites	**Datasets**	**Max**	**Mean**	**Min**	**Std**	**Max**	**Mean**	**Min**	**std**	**Max**	**Mean**	**Min**	**Std**
Córdoba	All	45.720	24.611	3.348	8.509	27.610	11.034	-8.300	6.270	9.693	3.672	0.355	2.350
	Train	45.720	24.417	3.348	8.541	25.160	11.031	-8.300	6.264	9.671	3.642	0.366	2.339
	Test	44.850	25.168	6.650	8.392	27.610	11.042	-3.431	6.289	9.693	3.758	0.364	2.377
Málaga	All	42.780	23.946	6.212	6.380	26.810	12.596	-4.270	5.565	10.309	3.439	0.453	1.954
	Train	42.780	23.851	6.212	6.411	25.200	12.546	-4.270	5.553	8.738	3.376	0.453	1.942
	Test	41.400	24.219	8.480	6.280	26.810	12.739	-1.185	5.596	10.309	3.620	0.501	1.977
Tabernas	All	42.550	23.244	4.282	7.275	26.020	9.794	-8.200	6.175	10.637	3.790	0.439	2.043
	Train	42.550	23.209	4.282	7.352	26.020	9.846	-8.200	6.173	10.637	3.760	0.439	2.048
	Test	41.700	23.349	5.274	7.034	23.660	9.639	-6.673	6.177	9.482	3.882	0.504	2.024
Conil de la Frontera	All	41.370	23.006	6.451	5.761	26.900	12.134	-5.383	5.375	9.353	3.213	0.434	1.876
	Train	41.370	23.100	6.451	5.770	25.970	12.080	-5.383	5.395	9.353	3.158	0.434	1.893
	Test	40.300	22.741	9.12	5.722	26.900	12.285	-1.298	5.316	9.189	3.369	0.436	1.820
Aroche	All	44.000	23.271	2.569	8.180	24.980	8.916	-8.000	5.655	8.751	3.267	0.311	2.086
	Train	44.000	23.206	4.430	8.143	24.980	9.100	-8.000	5.692	8.751	3.318	0.465	2.125
	Test	43.470	23.455	2.569	8.283	22.020	8.391	-5.113	5.516	7.324	3.121	0.311	1.963

The period of the datasets ranges from April 2000 to July 2018 for daily and semi-hourly data. The datasets were split into training and testing; the period from 2000 to 2014 was selected as training data, whereas the period from 2014 to 2018 was chosen as testing [50,60]. The reason why a hold-out technique is applied instead of cross-validation is that the dataset of this work is large, and composed of 18 years, so the risk of overfitting due to short datasets is low [62,63]. Apart from this split, a validation dataset was needed to check the performance of different hyperparameter sets in the process of searching the fittest tuning for each model - a random 15% from the training data was used for this purpose. The reason for taking a random period was to validate the performance in random years and days of the year; additionally, the seed of the random function was always the same, making all the validation dataset periods identical to one another. Once the fittest tuning (the most promising architecture and set of hyperparameter values on the training and validation datasets) was found, the whole training data was applied to train the model and, eventually, the testing dataset was used to obtain the performances.

2.2.Quality assurance procedures

In order to guarantee reliable results, the application of several quality control procedures to raw datasets was carried out. All this process is focused on identifying erroneous and questionable sensor measurements following the guidelines proposed by Estévez et al. [64] as a set of quality procedure checking tests (range test, internal consistency, step test, and persistence test) and the spatial consistency test (Estévez et al., 2018). The application of these quality assurance techniques has been successfully carried out under different climatic conditions worldwide (Estévez et al., 2016; Islam et al., 2019; Yi et al., 2018).

2.3.FAO56 – PM equation

Despite the differences between lysimeters measures and FAO-PM ET_0 estimates [65], the use FAO-PM ET_0 method is a common and accepted practice by numerous previous works under different climatic conditions [31,50,66–68]. Computations of ET_0 were carried out according to the standardized ASCE-FAO56-PM equation[1] and they were used as target values (Eq. 3.1):

$$\mathrm{ET_0} = \frac{0.408\Delta(\mathrm{Rn} - \mathrm{G}) + \gamma \frac{900}{\mathrm{Tm} + 273} \mathrm{u_2}(e_s - \mathrm{e}_a)}{\Delta + \gamma(1 + 0.34u_2)} \qquad \text{Eq. 3.1}$$

where ET_0 is the standardized grass reference evapotranspiration ($mm \cdot day^{-1}$), Δ is the slope of the saturation vapor pressure versus temperature curve ($kPa \cdot °C^{-1}$), Rn is the net radiation calculated at the crop surface ($MJ \cdot m^{-2} \cdot day^{-1}$), G is the soil heat flux density at the soil surface ($MJ \cdot m^{-2} \cdot day^{-1}$) which is assumed to be zero for daily values, Tm is the mean daily air temperature ($°C$), u_2 is the mean daily wind speed at 2m height ($m \cdot s^{-1}$), e_s and e_a are the saturation vapor pressure and the mean actual vapor pressure respectively (kPa), γ is the psychrometric constant ($kPa \cdot °C^{-1}$) and 0.408 is a coefficient ($MJ^{-1} \cdot m^2 \cdot mm$). The specification of the sensor used to calculate reference evapotranspiration in the assessed stations from RIA is shown in Table 3. 3.

Table 3. 3. RIA automated weather stations sensor specifications

Sensor device	Parameter	Accuracy	Range
Young 05103	Wind direction	±3%	From 0 to 360º
	Wind speed	±0.3 m/s	From 0 to 60 m/s
Pt1000	Temperature	±0.3 ºC	From -39.2 to 60ºC
Skye SP1110	Solar radiation	±5%	From 350 to 1100 nm
ARG 100	Precipitation	0.2 nm/tip	
Humicap 18’	Relative humidity	2% (from 0 to 90%)	From 0.8 to 100%
		3% (from 90 to 100%)	

2.4.Temperature-based methods

The main disadvantage of FAO PM, as previously mentioned, is that the required weather data are normally unavailable or of low quality on many sites worldwide. In this sense, Allen et al. [1] in the guidelines for FAO PM recommended two different methods: Hargreaves- Samani (HS) and the FAO56 computation using only temperature data (FAO5 PMT). HS is calculated according to equation 3.2,

$$ET_{0\,HS} = 0.0023\ (T_M + 17.8)\sqrt{T_X - T_N} \cdot R_a \qquad \text{Eq 3.2.}$$

where $ET_{0\ HS}$ is the ET_0 calculated through this method in $mm \cdot day^{-1}$; 0.0023 is an empirical coefficient; T_X, Tm, and Tn are the maximum, mean, and minimum air temperature (°C), respectively and R_a is the extraterrestrial radiation ($MJ{\cdot}m^{-2}{\cdot}day^{-1}$) calculated from the day of the year and the latitude.

In addition, this method can be locally calibrated to improve its performance. $ET_{0\,HSMOD}$ is mathematically expressed as equation 3,

$$ET_{0\,HSMOD} = m\ (T_M + n)\sqrt{T_X - T_N} \cdot R_a \qquad \text{Eq. 3.3}$$

where m and n are the calibration coefficients from Hargreaves and Samani [13], which need to be adjusted for site-specific conditions of the different stations using FAO PM as a reference, instead of using the general values of 0.0023 and 17.8, respectively.

Although other methods, such as the one proposed by Shiri et al. [69], could be useful to locally calibrate the coefficients of HS, the conventional procedure is adequate in cases where enough years are available on the dataset. On the other way, FAO PMT is the application of FAO56 PM using only temperature data and approximations to relative humidity (RH) and solar radiation by using air temperature, wind speed (u_2) is usually set to 2 ($m{\cdot}s^{-1}$) or regional mean values, the mean saturation vapor pressure (e_s), actual vapor pressure (e_a), and vapor pressure deficit (VPD) are estimated with equations 3.4, 3.5, and 3.6, respectively; dew temperature (T_{dew}) is estimated according to Todorovic et al. [21] where different approaches from Tn are carried out depending on the aridity (Table 3. 4).

$$e_s = \frac{0.6108}{2}\left(\exp\left(\frac{17.27Tx}{Tx+237.3}\right) + \exp\left(\frac{17.27Tn}{Tn+237.3}\right)\right)\ (\text{kPa}) \quad \text{Eq. 3.4}$$

$$e_a = 0.6108\exp\left(\frac{17.27T_{dew}}{T_{dew}+237.3}\right)\ (\text{kPa}) \quad \text{Eq. 3.5}$$

$$VPD = e_s - e_a\ (\text{kPa}) \quad \text{Eq. 3.6}$$

Table 3. 4. Correction of dew temperature (T_{dew}) estimation from minimum temperature (Tn) proposed by Todorovic et al. (2013).

Climate zones	**Annual Precipitation/ET0**	**Corrected T_{dew} (°C)**
Hyper arid	<0.005	T_{dew} = Tn – 4
Arid	0.05 – 0.20	T_{dew} = Tn – 2
Semi-arid	0.20 – 0.50	T_{dew} = Tn – 1
Dry sub-humid	0.50 – 0.65	T_{dew} = Tn – 1
Moist sub-humid	0.65 – 1.00	T_{dew} = Tn
Humid	>1.-0	T_{dew} = Tn

Finally, new parameters are introduced: $\text{Hourmin}_{\text{Sunset}}$ is the time of day in hours when the sunset occurs (calculated with longitude, latitude, and date), $\text{Hourmin}_{\text{Sunset}}$-$\text{Hourmin}_{\text{Tx}}$ ($Hourmin$) is the difference in hours between sunset and the time where T_x occurred (equation 3.7) and Energy_T is considered as the area below the temperature of a whole day (as it can be seen in Figure 3. 3) which is determined by equation 3.8, where T_i is the temperature every half hour in a day, from 00:00 to 23:30 (for example, T_0 is the temperature at 00:00, and T_1 is the temperature at 00:30).

$$Hourmin = Hourmin_{sunset} - Hourmin_{T_x}\ (\text{hours}) \quad \text{Eq. 3.7}$$

$$Energy_T = \sum_{i=1}^{48}(T_{i+1} - T_i) \cdot 0.5\ (°C \cdot hours) \quad \text{Eq. 3.8}$$

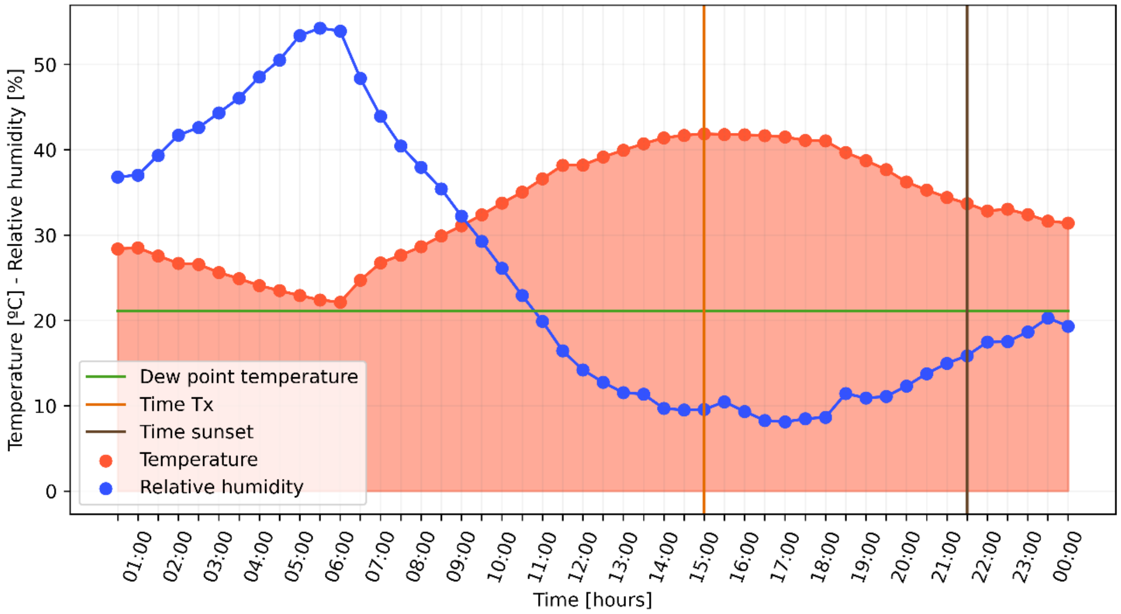


Figure 3. 3. Semihourly evolution of relative humidity and temperature, as well as dew point temperature, time Tx and time sunset in Córdoba on 7th July of 2015. The colored area below the temperature is the variable called EnergyT.

2.5.Machine learning models

All machine learning models were implemented on Python using the following libraries: Keras (Chollet et al,2015), Scikit-learn (Pedregosa et al, 2011), Scikit-optimize (Head et al., 2020, TensorFlow (Martín et al. 2015), XGBoost (Chen and Guestrin, 2016), NeuPy and hpelm (Akusok et al., 2015). All computations were running in a server workstation over Anaconda 3.7 with the following features: 2x Intel® Xeon® CPU E5-2650 v3 @ 2.30GHz and 128 GB of RAM. The models (ANN, SVM, RF, XGBoost, GRNN, and ELM) used in this study are detailed below, as well as the hyperparameters optimization in section 2.5.7.

2.5.1. Multilayer Perceptron (MLP)

A Neural Network (NN) model is a data processing system inspired by the biological nervous system of the human brain. It is composed of a large number of interconnected neurons working on the purpose of solving a specific problem; in our case, for estimating reference evapotranspiration using only temperature parameters as inputs.

All NN models have three different layers: the input, the hidden, and the output layer. The input layer corresponds to the data our system can use to resolve a problem, whereas the output represents the results. The hidden layer is usually called as a black box because its processes are not controlled, and its functionality is modified by an auto-training method. A single hidden layer and a single neuron model is called perceptron (Figure 3. 4). Equation 3.9 represents the mathematical expression of a general perceptron with different inputs, where x_i represents the different inputs (i) of our system, w_i are the weights that multiply each input to a corresponding neuron, b is a bias value, f is an activation function and y is the output.

$$y = f\left(\sum_{i=1}^{m} w_i * x_i + b\right) \qquad \text{Eq. 3.9}$$

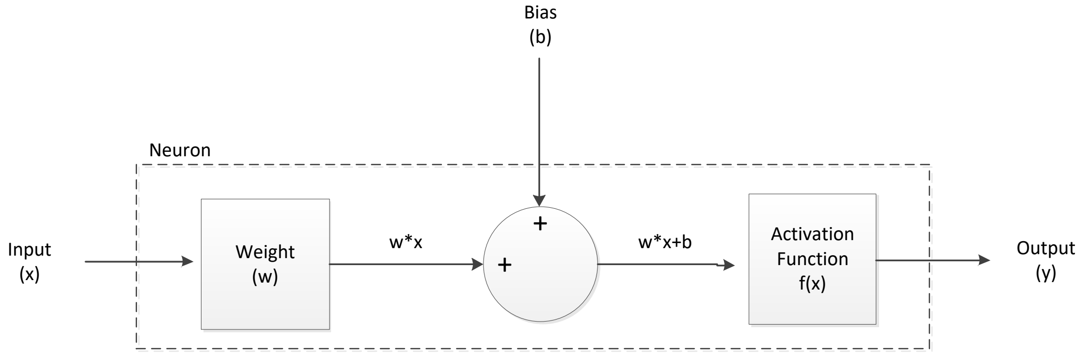


Figure 3. 4. One neuron structure

Besides, multiple hidden layers with one or more neurons each are called Multilayer Perceptron (MLP) and their architecture is represented in the appendix (Figure 3. 5). The f is responsible for whether a neuron should be activated or not. The activation functions used in this work have been: sigmoid, tanh, and ReLU (Figure 3. 6).

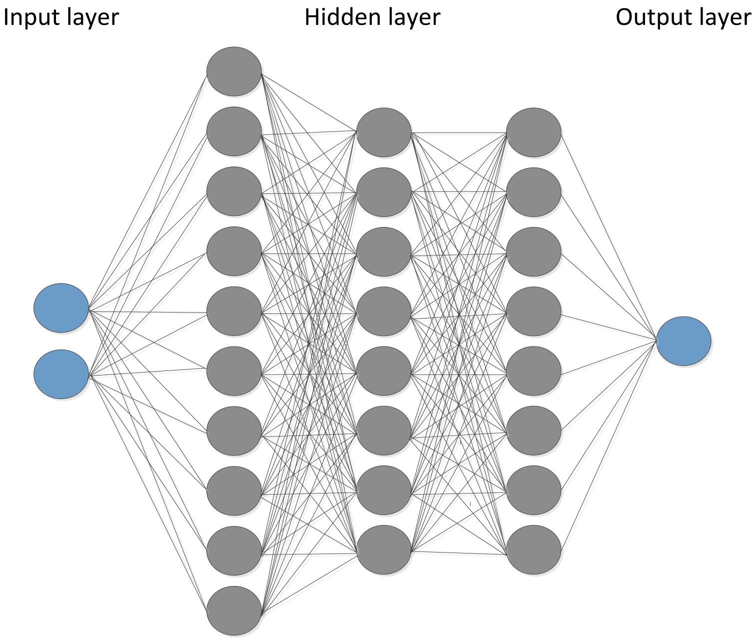


Figure 3. 5. Multilayer Perceptron (MLP) architecture

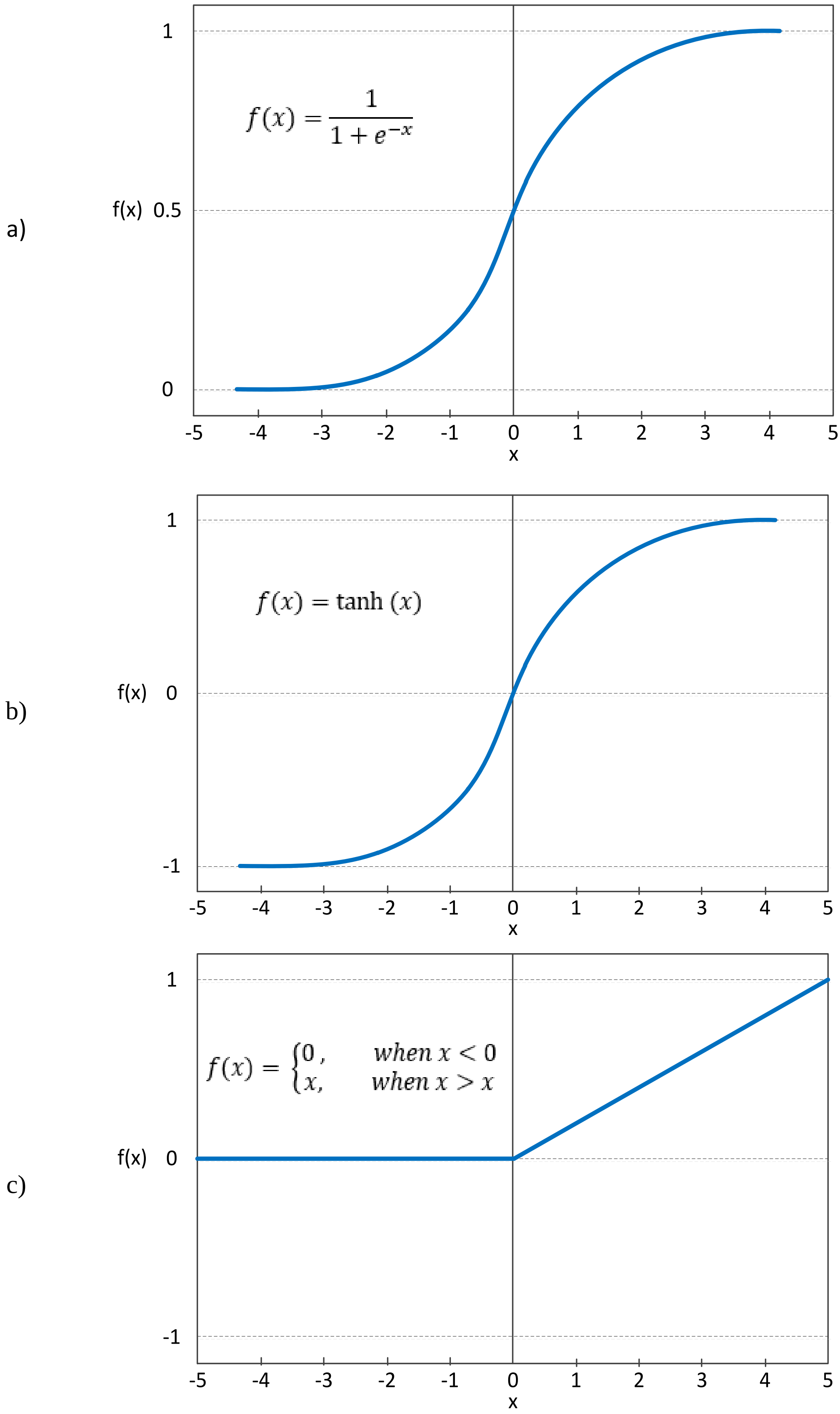


Figure 3. 6. Activation functions. a) sigmoid, b) tanh, c) ReLU.

The most important part of MLP is the learning process, which is called back-propagation, where the results given by the model are introduced in a cost function and the weights are updated every epoch to minimize it. This updating process can be carried out by several optimizers. In this study, the optimizers SGD (Stochastic Gradient Descent), RMSprop (Root Mean Square prop), and Adam (Adaptive Moment Estimation) have been used. The SGD optimizer is the most used due to its simplicity and good results. RMSprop optimizer is like gradient descent with momentum; the difference lies in how the gradients are calculated. Eventually, the Adam is a combination of RMSprop and SGD Descent with momentum, using the squared gradients to scale the learning rate like RMSprop, and taking the momentum by using the moving average of the gradient. For more detailed information about these optimizers, [70–72] can be consulted.

Finally, the hyperparameters introduced into optimization are the number of neurons of the first hidden layer (from 1 to 20 neurons), the number of neurons of the other hidden layers (from 1 to 20), the number of hidden layers (from 1 to 4), the activation (sigmoid, tanh, and ReLU), the optimizer function (SGD, RMSprop, and Adam) and the number of epochs (from 1 to 100).

2.5.2. Extreme Learning Machine (ELM)

Despite the good results given by MLP in reference to evapotranspiration, the high computational cost of the learning process made scientists study new approaches. Huang et al. [73] proposed a single hidden layer feedforward neural network (SLFNNs), where the weights and biases of the hidden layer are randomly generated and not tuned, and the output weights and bias are analytically calculated. As a result, the model obtained has a very low computational cost because no iteration learning process is required.

Even if the number of neurons is less than the number of inputs, the hidden node parameters of ELM should not be tuned throughout training, being able to learn distinct samples with good results [42,46,48]

The hyperparameters introduced into optimization are the number of neurons of the hidden layer from 1 to 1000 and its activation function.

2.5.3. Generalized Regression Neural Network (GRNN)

Generalized Regression Neural Networks (GRNN) were first proposed by Specht [74], as a modification of a Radial Basis Function Network (RBFN). GRNN models are based on nonlinear regression estimation functions and their structure is composed of four layers: the input layer, the pattern layer, the summation layer, and the output layer (Figure 3. 7). Each neuron of the pattern layer represents a training input pattern and its output is a measure of the distance of the input from the stored patterns. The summation layer is composed of two different kinds of nodes, the S-nodes, and the D-node, represented by S and D in Figure 3. 7. The number of S-nodes is the same as the number of outputs and it computes the sum of the weighted outputs of the pattern layer; each node is connected to its correspondent output. On the contrary, the D-node is connected to all outputs and it calculates the unweighted outputs of the pattern layer. The calculation of each output is the division between the output of each S-node by the D-node output, yielding the estimated value (Figure 3. 7). The mathematical expression of the output is represented in Equations 3.10 and 3.11:

$$y_i = \frac{\sum_{i=1}^{N} w_i * \exp\left(-G(x, x_i)\right)}{\sum_{i=1}^{N} \exp\left(-G(x, x_i')\right)} \tag*{Eq. 3.10}$$

$$G(x,\ x_i') = \sum_{j=1}^{M} \left(\frac{x - x_j}{\sigma}\right)^2 \tag*{Eq. 3.11}$$

where y_i is the predicted value to an unknown input vector x, x_i is the input training vector, w_i is the weight connection between the ith neuron, N and M are the numbers of training patterns and the number of elements of an input vector respectively, G is the Gaussian function, and σ is a hyperparameter that modifies the width of the Gaussian function [75].

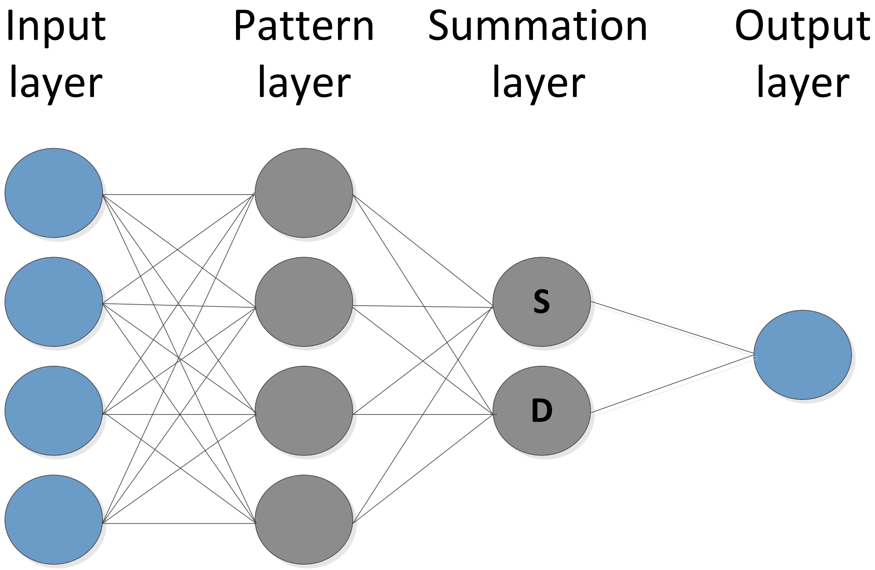
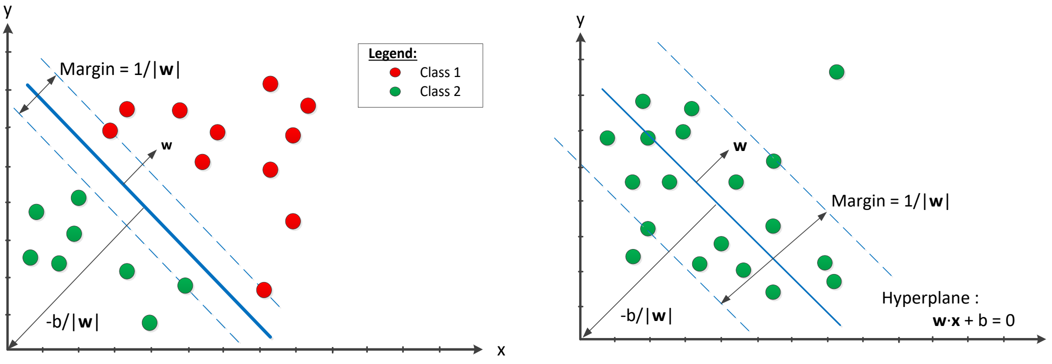


Figure 3. 7. GRNN architecture for two inputs and two outputs.

The special feature of GRNNs is that it does not require an iterate training procedure because the local minima problem was not faced during development [74], making this model the fastest and with a low error performance [42,43,47]. This algorithm is designed by '*NeuPy*' python library, where the only hyperparameter to modify is σ with a range of [0.0001-5.0].

2.5.4. Support Vector Machine (SVM)

Support Vector Machine is a machine learning technique that covers classification (SVM) and regression (SVR) alike. The main idea of SVM, which is extrapolated to SVR, is to linearly separate two or more classes with a hyperplane, where the margins are maximized. In Figure 3. 8 there is a graphical summary of Support Vectors for classification (a) and regression (b).



a) b)

Figure 3. 8. Support vector machines for classification tasks (SVM) and regression (SVR). a) SVM, b) SVR.

The optimization problem is mathematically expressed as equation 3.12 or by Lagrange function as equations 3.13 and 3.14.

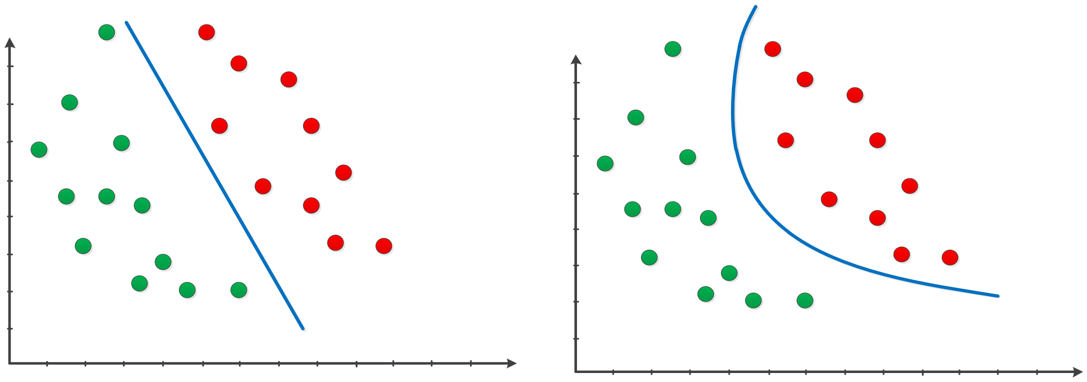
$$min\frac{1}{2}\|w\|^2 \qquad s.t. \qquad y_i(\boldsymbol{x_i}\cdot\boldsymbol{w}+b)-1\ \geq 0 \qquad \forall i \qquad \text{Eq. 3.12}$$

$$L(\boldsymbol{w},b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i[\,y_i(\boldsymbol{x_i}\cdot\boldsymbol{w}+b)-1\,] \qquad \text{Eq. 3.13}$$

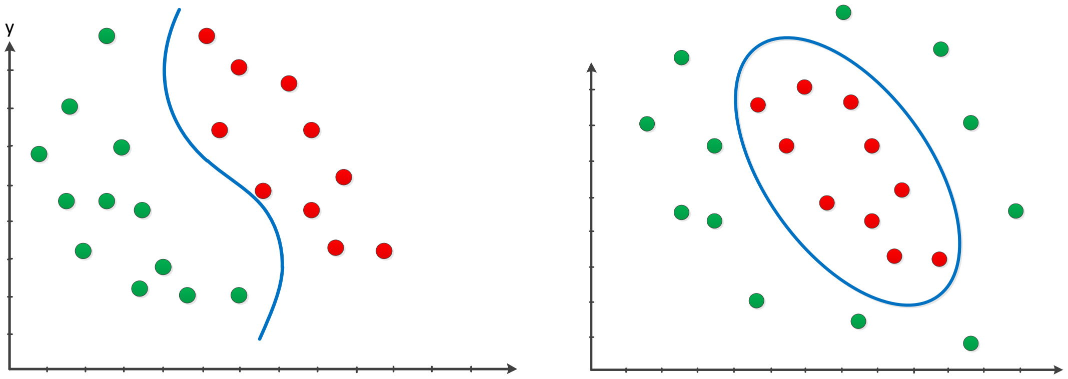
$$\min_{\boldsymbol{w},b}\max_{\alpha}\ L(\boldsymbol{w},b,\alpha) \qquad \text{s.t.} \qquad \alpha_i\geq 0, i=1,\ldots,m \qquad \text{Eq. 3.14}$$

where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x_i}$ is the input vector, b is the bias, y_i is the actual class/value of the input and α is the Lagrange multiplier.

An important aspect is the use of Kernel functions to enable support vectors to operate in a high-dimensional feature (**x**) space. The definition is: “*Given a mapping function* $\emptyset: X\rightarrow V$, *we call the function* $K: X\rightarrow\ \mathbb{R}$ *defined by* $K(\boldsymbol{x},\boldsymbol{x}')=\big(\emptyset(\boldsymbol{x}),\emptyset(\boldsymbol{x}')\big)v$, *where* $(\cdot,\cdot)v$ *denotes an inner product in* V, *a kernel function*”(Vapnik, 2000). The different kernels we can find in *Scikit-learn* are linear kernel $K(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{x}\cdot\boldsymbol{x}'$, polynomial kernel $K(\boldsymbol{x},\boldsymbol{x}')=(\boldsymbol{x}\cdot\boldsymbol{x}'+c)^d$, sigmoid kernel $K(\boldsymbol{x},\boldsymbol{x}')=tanh(\boldsymbol{x}\cdot\boldsymbol{x}'+c)$ and RBF or gaussian kernel $K(\boldsymbol{x},\boldsymbol{x}')=\exp(-\gamma\|\boldsymbol{x}\cdot\boldsymbol{x}'\|^{\mathbf{2}})$. Figure 3. 9 shows an example for each kernel. For further SVR explanations, see Smola and Schölkopf [76].



a) b)



c) d)

Figure 3. 9. Kernel application types. a) Linear kernel, b) Polynomial kernel, c) Sigmoid kernel and d) RBF kernel.

The hyperparameter used in optimization is the kernel function (polynomial, sigmoid, and RBF), c (from 0.0001 to 1000), and ε (from 0.0001 to 1000). The parameters c and ε correspond to the regularization parameter and the margin within no penalty is associated, respectively.

2.5.5. Random Forest (RF)

Random Forest network is an ensemble tree-based method proposed by Breiman [77] where trees are trained in a random subset and it can be used with bagging or pasting, although is more common to use the first technique. It also introduces extra randomness when the number of trees starts growing. Instead of searching for the best feature to split a node, it searches for the best feature among a random subset of features, resulting in greater tree diversity. For further information, Feng et al. [42] can be consulted.

The hyperparameters tested in this work are the number of trees in the forest (from 1 to 1000), the splitting criteria (the mean square error o mean absolute error), the maximum number of features considered to split a node of a tree (one third, one half, three fourth, the square root and the logarithm in base two of the number of features) and the use of bagging or pasting. The rest of the parameters are in default mode.

2.5.6. XGBoost

It is a new machine learning model proposed by Chen et al. [78], based on Gradient Boosting Machines (GBM) and the Classification And Regression Trees (CART) where a subset of weak learners are combined to develop new stronger learners through the introduction of training strategies - the main idea of boosting methods is to train models sequentially, where each model try to reduce the error of the predecessor. The objective of XGBoost is to reduce the computational cost while preventing overfitting, thanks to the capacity of executing parallel calculations.

The hyperparameters introduced in the optimization process are the number of trees in the forest (from 1 to 1000), the learning rate (from 0.0001 to 1), the maximum depth of a tree (from 1 to 30), the subsample ratio of the training instances, setting it to 0.5 means that XGBoost would randomly sample half of the training data before growing trees, and this will prevent overfitting (from 0.01 to 0.999), the subsample ratio of columns when constructing each tree (1/3, 1/2, 3/4 and 1) and the objective function (regression with squared loss and regression with squared log loss).

2.5.7. Bayesian hyperparameter optimization

The aim of hyperparameter optimization inside machine learning consists of finding a hyperparameter set for a given machine learning model that returns the most accurate performance on the validation dataset, and likely, it will outperform in the testing dataset. For example, for a typical multilayer perceptron (MLP) model, it is very common to search for the number of hidden layers, the number of neurons in those layers, the learning rate, or the optimizer function, among others.

Hyperparameter optimization is mathematically represented as follows (Equation 3.15):

$$x^{*} = \arg\min f(x), \quad x \in X \qquad \text{Eq. 3.15}$$

where $f(x)$ represents the objective function we want to minimize (for example, the mean absolute error between predicted values and measured values), x^{*} is the optimized set of hyperparameters and X represents the whole hyperparameter space.

The first approach is to manually change parameters based on experience and previous results until a good set is found, known as manual tuning. Sometimes it could yield good results, although not the best, but in many other cases, it could lead to a local minimum which could be far from the global. As a second solution, automatic methods such as grid or random search could be used, having slightly better results than manual tuning because all this process is automatically carried out in a loop. However, they are completely blind to past evaluations, and as a result, they often spend a significant amount of time evaluating bad parameters. Bayesian optimization emerged as an efficient solution to these problems [79]. It is an algorithm that uses the Bayes theorem to search the minimum of an objective function. The procedure is: 1) a defined number of points from the hyperparameter space are tested; 2) the algorithm considers past evaluations when selecting the next hyperparameters to evaluate (acquisition function), enabling itself to focus on those areas of the parameter space that are believed to bring promising validation scores; 3) this new hyperparameter set is evaluated and 4) if the optimization has not finished, it goes to the second point.

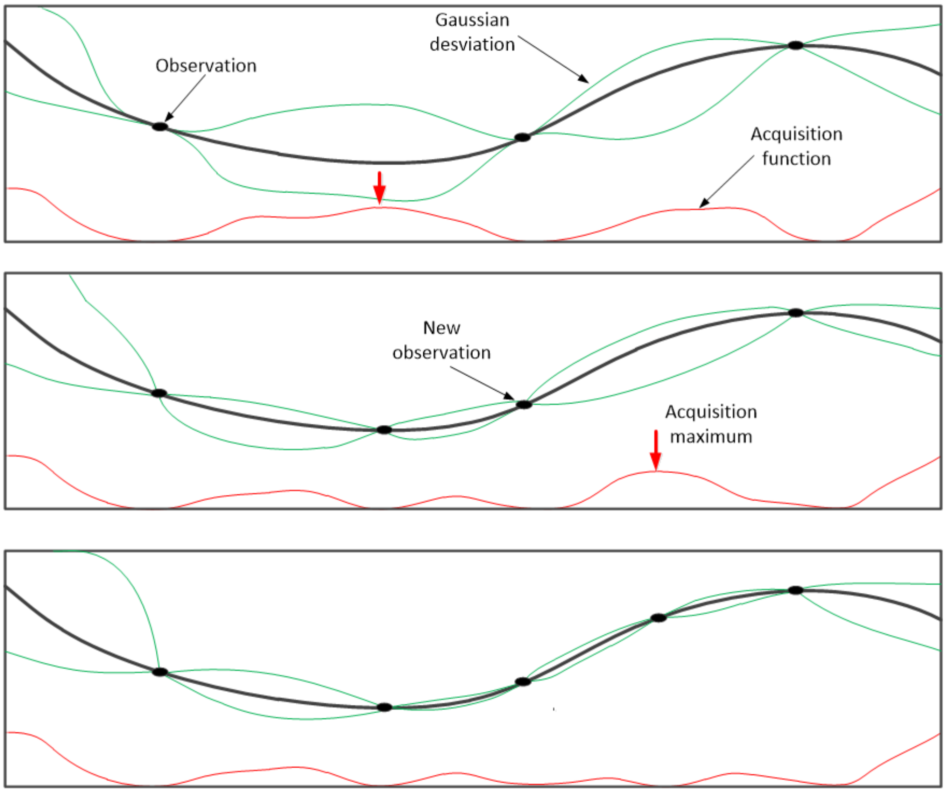


Figure 3. 10. An example of Bayesian optimization.

Roughly, in terms of machine learning, it could be defined as the automatic method of manual tuning due to its similarity to human behavior while tuning [80,81]. Three major choices must be made to select a prior over functions that will express assumptions about the function being optimized (in our case, the Gaussian Process (GP) prior due to its flexibility): the acquisition function, determining the posterior model, and allowing us to determine the next point to evaluate [82]. Figure 3. 10 shows an example of the Bayesian optimization process for GP. Finally, it is necessary to select which hyperparameters are used and their limits. In this aspect, several ranges of hyperparameters were chosen by trial and error [82]. It is important to realize that the bigger the limits, the bigger the number of epochs the algorithm needs to get the most accurate set of parameters, and the longer the computational cost – this is the reason why choosing a sensible range is very important to the efficiency of the optimization.

Besides, this methodology should be carried out using the training dataset, whereas the final neural network architecture will be assessed by the testing dataset (these data cannot influence the training or tuning processes), which is in line with works like [36,42,47,50] However, using the testing dataset to tune the fittest architecture and a hyperparameter set is not recommended because it conducts the results to a bias on testing data, which alienates from the main goal of generalization.

The specific limits and hyperparameters were previously defined for each model, whereas the number of epochs for the optimization was set to 500, being 50 epochs randomly selected. For further information about Bayesian optimization and Gaussian processes, see Shahriari et al. [81].

2.6.Data standardization

Standardization of input datasets is a common requirement among all neural network models, in the exception of tree-based models, and it can be expressed as equation 3.16:

$$x_{stand} = \frac{x - \bar{x}}{\sigma} \qquad \text{Eq. 3.16}$$

where x_{stand} is the data standardized, x is the input data, $\bar{x}$ is the mean of the training input dataset and σ is the standard deviation of the training input dataset.

2.7.Statistical analysis

The model performances were evaluated by using the following parameters: Root Mean Square Error ($RMSE$), Mean Bias Error (MBE), the coefficient of determination (R^2) and the Nash-Sutcliffe model efficiency coefficient (NSE). The RMSE, MBE, $R^{2,}$ and NSE are defined as equations 3.17, 3.18, 3.19, and 3.20:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2} \qquad \text{Eq. 3.17}$$

$$MBE = \frac{1}{m}\sum_{i=1}^{m} meas_i - pred_i \qquad \text{Eq. 3.18}$$

$$R^2 = \frac{\left(\sum_{i=1}^{m}(meas_i - \overline{meas})\left(pred_i - \overline{pred}\right)\right)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2 \sum_{i=1}^{m}\left(pred_i - \overline{pred}\right)^2} \qquad \text{Eq. 3.19}$$

$$NSE = 1 - \frac{\sum_{i=1}^{m}(meas_i - pred_i)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2} \qquad \text{Eq. 3.20}$$

where m is the number of records, the bar denotes the mean, and pred and meas are the predicted and FAO56-PM ET_0 values, respectively.

3. Results and Discussion

In order to avoid not promising configurations, a previous evaluation was carried out in the weather station of Córdoba with different combinations of T_X, Tn, R_a, $Energy_T$, e_s, e_a, VPD, T_{dew} and Hourmin, where 11 configurations from 3 to 6 inputs were chosen (Tx, Tn, and Ra were always selected in all the configurations due to their high efficiency). The rest of the combinations were excluded because no significant improvements were shown. Finally, these selected 11 input configurations (Table 3. 5) were assessed at each location and for all models.

Table 3. 5. Input climatic parameter configurations (T_x: Maximum temperature, T_n: Minimum temperature, T_{dew}: dew point temperature, R_a: Extraterrestrial radiation, e_s: mean saturation vapor pressure, e_a: actual

vapor pressure, VPD: vapor pressure deficit, $Energy_T$: the integral of the daily thermal signal, Hourmin is the difference between the time when sunset occurs and the time the maximum temperature occurs).

Configurations	**Climatic parameters**	**Number of inputs**
1	T_X, Tn, Ra	3
2	T_X, Tn, Ra, $Energy_T$	4
3	T_X, Tn, Ra, Hourmin	4
4	T_X, Tn, Ra, e_s	4
5	T_X, Tn, Ra, e_a	4
6	T_X, Tn, Ra, VPD	4
7	T_X, Tn, Ra, T_{dew}	4
8	T_X, Tn, Ra, $Energy_T$, VPD	5
9	T_X, Tn, Ra, e_a, e_s	5
10	T_X, Tn, Ra, e_a, T_{dew}	5
11	T_X, Tn, Ra, $Energy_T$, VPD, T_{dew}	6

3.1.Performance of models in the different locations

In terms of the statistics MBE, RMSE, $R^{2,}$ and NSE, their minimum, mean and maximum values are summarized in Table 3. 6 for all the configurations of the different studied models at each station.

In Cordoba (aridity index = 0.4616), in terms of RMSE, the mean values of all ML models (0.5249mm/day) improved the performance of HSMOD (0.6071 mm/day) and HS (0.6460 mm/day). The best RMSE performance (0.4983 mm/day) was carried out by SVM, whereas the worst among ML models was carried out by GRNN (0.5691 mm/day), so close to XGBoost (0.5620 mm/day). On average, the best model performance was SVM with the mean RMSE value of 0.5068 mm/day, very close to ELM (0.5145 mm/day). Regarding NSE, its behavior was like RMSE, where all ML performances (the worst value was 0.9427, obtained by GRNN) outperformed HSMOD (0.9348). The most precise model was, again, SVM (0.9561) followed by MLP (0.9557) and ELM (0.9545); the ranking on mean values is almost identical, where SVM maintain the first place, but MLP and ELM switch their positions with the values 0.9546, 0.9532 and 0.9510, respectively. The minimum MBE value (-0.0006 mm/day) of all the models assessed was obtained with MLP (it also got the worst value among ML models, 0.1931 mm/day), although on average, the best performance was carried out by SVM with a mean of 0.0073 mm/day. The R^2 values were very satisfactory for all models, ranging from 0.9434 (GRNN) to 0.9565 (MLP), even in HS (0.9366) and HSMOD (0.9358).

Table 3. 6. Minimum (Min), mean, and maximum (Max) values of MBE, RMSE, R^2, and NSE for all the stations using different models: Hargreaves-Samani (HS), Calibrated Hargreaves-Samani (HSMOD), Multilayer Perceptron (MLP), Extreme Learning Machines (ELM), Support Vector Machines (SVM), Generalised Regression Neural Network (GRNN), XGBoost and Random Forest (RF). The symbol * means ‘*no-value*’.

Station	Model	MBE (mm/day) Min/Mean/Max	RMSE (mm/day) Min/Mean/Max	R^2 Min/Mean/Max	NSE Min/Mean/Max
Córdoba	HS	*/0.2339/*	*/0.6460/*	*/0.9366/*	*/0.9261/*
	HSMOD	*/0.0690/*	*/0.6071/*	*/0.9358/*	*/0.9348/*
	MLP	-0.0006/0.0458/0.1931	0.5005/0.5261/0.5462	0.9534/0.9547/0.9565	0.9472/0.9510/0.9557
	ELM	0.0315/0.0409/0.0453	0.5069/0.5145/0.5250	0.9523/0.9541/0.9556	0.9512/0.9532/0.9545
	SVM	0.0044/0.0073/0.0135	0.4983/0.5068/0.5134	0.9536/0.9548/0.9564	0.9534/0.9546/0.9561
	GRNN	0.0444/0.0511/0.0547	0.5208/0.5351/0.5691	0.9434/0.9505/0.9531	0.9427/0.9493/0.9520
	XGBoost	0.0082/0.0156/0.0457	0.5178/0.5307/0.5620	0.9446/0.9508/0.9531	0.9441/0.9502/0.9526
	RF	0.0407/0.0466/0.0509	0.5217/0.5363/0.5466	0.9482/0.9502/0.9529	0.9472/0.9491/0.9519
Málaga	HS	*/-0.0630/*	*/0.6876/*	*/0.8844/*	*/0.8791/*
	HSMOD	*/-0.1337/*	*/0.6781/*	*/0.8874/*	*/0.8824/*
	MLP	-0.0732/-0.1433/-0.2177	0.5952/0.6189/0.6410	0.9070/0.9081/0.9105	0.8949/0.9019/0.9091
	ELM	-0.1142/-0.1230/-0.1291	0.6111/0.6173/0.6256	0.9038/0.9064/0.9076	0.8999/0.9025/0.9042
	SVM	-0.1381/-0.1564/-0.1774	0.6189/0.6243/0.6349	0.9050/0.9067/0.9084	0.8969/0.9003/0.9017
	GRNN	-0.1213/-0.1286/-0.1361	0.6186/0.6254/0.6417	0.8988/0.9042/0.9061	0.8947/0.8999/0.9019
	XGBoost	-0.1264/-0.1603/-0.1760	0.6300/0.6438/0.6512	0.8987/0.9005/0.9034	0.8916/0.8939/0.8982
	RF	-0.1238/-0.1333/-0.1395	0.6184/0.6295/0.6388	0.9004/0.9033/0.9062	0.8957/0.8986/0.9019
Tabernas	HS	*/-0.1836/*	*/0.7302/*	*/0.8782/*	*/0.8698/*
	HSMOD	*/-0.0728/*	*/0.7011/*	*/0.8817/*	*/0.8800/*
	MLP	0.0015/-0.0634/-0.2737	0.6273/0.6604/0.7109	0.8938/0.8987/0.9075	0.8767/0.8935/0.9040

	ELM	-0.0427/-0.0557/-0.0654	0.6210/0.6532/0.6722	0.8907/0.8967/0.9064	0.8897/0.8958/0.9059
	SVM	-0.1010/-0.1271/-0.1387	0.6349/0.6744/0.6874	0.8908/0.8945/0.9046	0.8847/0.8889/0.9016
	GRNN	-0.0644/-0.0680/-0.0750	0.6583/0.6690/0.6877	0.8857/0.8920/0.8956	0.8846/0.8908/0.8943
	XGBoost	-0.0846/-0.1017/-0.1121	0.6595/0.6724/0.6901	0.8868/0.8926/0.8964	0.8838/0.8897/0.8939
	RF	-0.0540/-0.0615/-0.0668	0.6518/0.6668/0.6794	0.8883/0.8924/0.8972	0.8874/0.8915/0.8963
Conil	HS	*/-0.0683 /*	*/0.8096/*	*/0.8064/*	*/0.8022/*
	HSMOD	*/-0.2830/*	*/0.8264/*	*/0.8182/*	*/0.7939/*
	MLP	-0.0048/-0.1702/-0.3117	0.5853/0.6072/0.6619	0.8988/0.9005/0.9032	0.8679/0.8887/0.8967
	ELM	-0.1668/-0.1786/-0.1967	0.5985/0.6204/0.6721	0.8745/0.8938/0.9005	0.8637/0.8838/0.8920
	SVM	-0.2153/-0.2407/-0.2592	0.6208/0.6384/0.6444	0.8921/0.8948/0.8978	0.8748/0.8771/0.8838
	GRNN	-0.1511/-0.1588/-0.1665	0.5884/0.5969/0.6137	0.8935/0.9002/0.9040	0.8864/0.8925/0.8956
	XGBoost	-0.2202/-0.2256/-0.2333	0.6284/0.6382/0.6581	0.8862/0.8927/0.8964	0.8694/0.8771/0.8809
	RF	-0.1560/-0.1600/-0.1673	0.5859/0.6069/0.6177	0.8927/0.8969/0.9039	0.8849/0.8889/0.8965
Aroche	HS	*/0.7430/*	*/0.9243/*	*/0.9472/*	*/0.7783/*
	HSMOD	*/0.3347/*	*/0.5832/*	*/0.9466/*	*/0.9117/*
	MLP	0.1789/0.3038/0.4185	0.4592/0.5323/0.6187	0.9561/0.9579/0.9599	0.9007/0.9259/0.9453
	ELM	0.2577/0.2845/0.2985	0.4800/0.5179/0.5362	0.9545/0.9574/0.9617	0.9254/0.9303/0.9402
	SVM	0.2469/0.2703/0.2857	0.4829/0.5134/0.5386	0.9562/0.9580/0.9609	0.9247/0.9315/0.9395
	GRNN	0.2826/0.2959/0.3053	0.5185/0.5402/0.5592	0.9497/0.9532/0.9568	0.9189/0.9243/0.9302
	XGBoost	0.2403/0.2616/0.2786	0.4924/0.5168/0.5457	0.9482/0.9545/0.9581	0.9227/0.9306/0.9371
	RF	0.2759/0.2880/0.2987	0.5170/0.5347/0.5468	0.9511/0.9528/0.9565	0.9224/0.9258/0.9307

In Málaga (aridity index = 0.3666), in terms of NSE, the efficiency performance obtained by all the studied models outperformed HS (0.8791) and HSMOD (0.8824). Their values ranged from 0.8916 (XGBoost) to 0.9091 (MLP), and the mean among all ML models was 0.8995. MLP, ELM, and SVM overpassed the 0.9 performance in mean, whereas all models, except for XGBoost, exceeded it with their best configuration. Besides, regarding RMSE, the values were also better than those obtained by Hargreaves methods, being the best mean values obtained by ELM (0.6173 mm/day), MLP (0.6189 mm/day), and SVM (0.6243 mm/day), in this order. MLP outperformed the rest of the models with the best RMSE value (0.5952 mm/day). On the other hand, XGBoost got the worst performance with an RMSE value of 0.6512 mm/day, which keeps being better than those obtained by HS (0.6876 mm/day) or HSMOD (0.6781 mm/day). Regarding R^2, the results of ML models outperformed HS and HSMOD. However, the MBE values were worse on average, with exception of ELM, GRNN, and RF. The best value was obtained, in both cases, by MLP (MBE=-0.0732 mm/day and R^2= 0.9105).

In Tabernas, which is the station with the lowest aridity index (0.178), the RMSE ranged from 0.6210 mm/day (ELM) to 0.7109 mm/day (MLP). This last model and configuration was the only one that did not outperform HSMOD (RMSE=0.7011 mm). From all the models assessed, ELM was the one with the smallest minimum, mean, and maximum RMSE values (0.6210, 0.6532, and 0.6722 mm/day respectively). Concerning NSE, its values ranged from 0.8767 (MLP) to 0.9059 (ELM), being 0.8767 the unique value that did not outperform HSMOD (0.8800). On average, the model performance raking in both, NSE and RMSE, is ELM, MLP, RF, GRNN, XGBoost, and SVM, in this order. In addition, the MBE value reached its minimum 0.0015 mm/day with MLP and configuration 9 and, on average, the value obtained using HSMOD (-0.0728 mm) was outperformed by all the models, except for SVM and XGBoost. Finally, the R^2 obtained a very similar behavior to NSE, but in this case, all ML models outperformed HS and HSMOD.

In Conil (aridity index = 0.4790), the best models regarding RMSE and NSE were MLP (RMSE=0.5853 mm/day and NSE=0.8967), RF (RMSE=0.5859 mm/day and NSE=0.8965) and GRNN (RMSE=0.5884 mm/day and NSE=0.8956). In contrast, the worst model performances were ELM (RMSE = 0.6721 mm/day and NSE=0.8637), MLP (RMSE=0.6619 mm/day and NSE 0.8679), and XGBoost (RMSE=0.6581 mm/day and NSE=0.8694) in this coastal location. Moreover, the ML models obtained the most significant improvements against HSMOD and HS over the other stations. The R^2 behavior was in line with NSE, but the MBE of ML models did not outperform HS (-0.0683 mm/day), which was one of the best apart from MLP (-0.0048 mm/day). A main feature of MBE is that all results were negative, which means that all models overestimated ET_0.

Finally, in Aroche, which is the most humid location (aridity index=0.555) the values of R^2 and NSE using HSMOD and all evaluated models were very satisfactory, being both statistics higher than 0.9 for all the cases. According to RMSE, the best results were obtained by MLP (0.4592 mm/day) followed by ELM (0.4800 mm/day). However, on average, SVM had the highest accurate estimations (0.5134 mm/day), followed very close to XGBoost and ELM (0.5168 and 0.5179 mm/day, respectively); being all of them lower than the obtained by HSMOD. On the other hand, MBE denotes underestimations in all cases, ranging from 0.1789 mm/day with MLP to 0.4185 mm/day using MLP.

Comparing the performance of the different models in the different stations it could be stated that there was a high-performance variability between models in the different stations. MLP was usually one of the best models in most of the statistics, however, it usually also had one of the worst values; there is a higher dependence on the chosen configuration, which could lead to overfitting in some cases. The ELM and SVM usually obtained very close or even better values than MLP, apart from having a less scattered performance variance. On the other hand, RF and XGBoost (tree-based models) were in general the worst models in the different statistics. The GRNN and ELM are the models with less computational requirements because they do not need any training process. It could be stated that ELM is the most adequate model to estimate ET_0 in these weather stations

Comparing results from different stations (with very close mean annual ET_0 values – table 2), Córdoba and Tabernas (inland locations) had very accurate MBE results, while the coastal stations denoted an overestimation tendency in all models and configurations. The RMSE values obtained in Málaga, Tabernas, and Conil stations were quite similar on average, although Tabernas had slightly worse values in all the different models. On the other hand, Córdoba and Aroche had the best results. In terms of R^2 and NSE, the results were in line with RMSE: Córdoba and Aroche had the best estimations with values around 0.95 in both locations, but the rest of the stations were closer to 0.9, which keep being considered as a good performance.

3.2.Performance of the configurations

In order to evaluate the results obtained between different configurations in each station and without any dependence on ML models, the mean of the statistical indices RMSE, MBE, $R^{2,}$ and NSE of the different models using a specific configuration are shown in Figure 3. 11. In general, the input configuration T_X, Tn, and R_a (configuration 1) performed RMSE as the worst value in all the stations. On the other hand, the statistics MBE, $R^{2,}$ and NSE went along with this line in inland locations (Córdoba, Tabernas, and Aroche), whereas in coastal stations (Málaga and Conil), not such bad performances were obtained using this configuration. In some cases, very satisfactory values were obtained in coastal locations, such as MBE in Conil and Málaga, and R^2 and NSE in Conil. The introduction of the $Energy_T$ parameter in the inputs (configurations 2, 8, and 11) improved the performances in all the stations, except for MBE in Conil. In Málaga, the lowest RMSE and the highest NSE values were obtained using configurations 11 followed by 2 and 8. Moreover, in Conil, the results obtained using configurations 2, 7, and 6 were the best in terms of RMSE and NSE. Córdoba and Aroche had, on average, the best R^2 performance and the lowest variability using most configurations, which is in contrast to the rest of the stations. However, the NSE values obtained in Córdoba surpassed the values obtained in Aroche for all configurations. On the other hand, Aroche had the highest MBE values, which is the indication of a dominant underestimation tendency for all the configurations assessed.

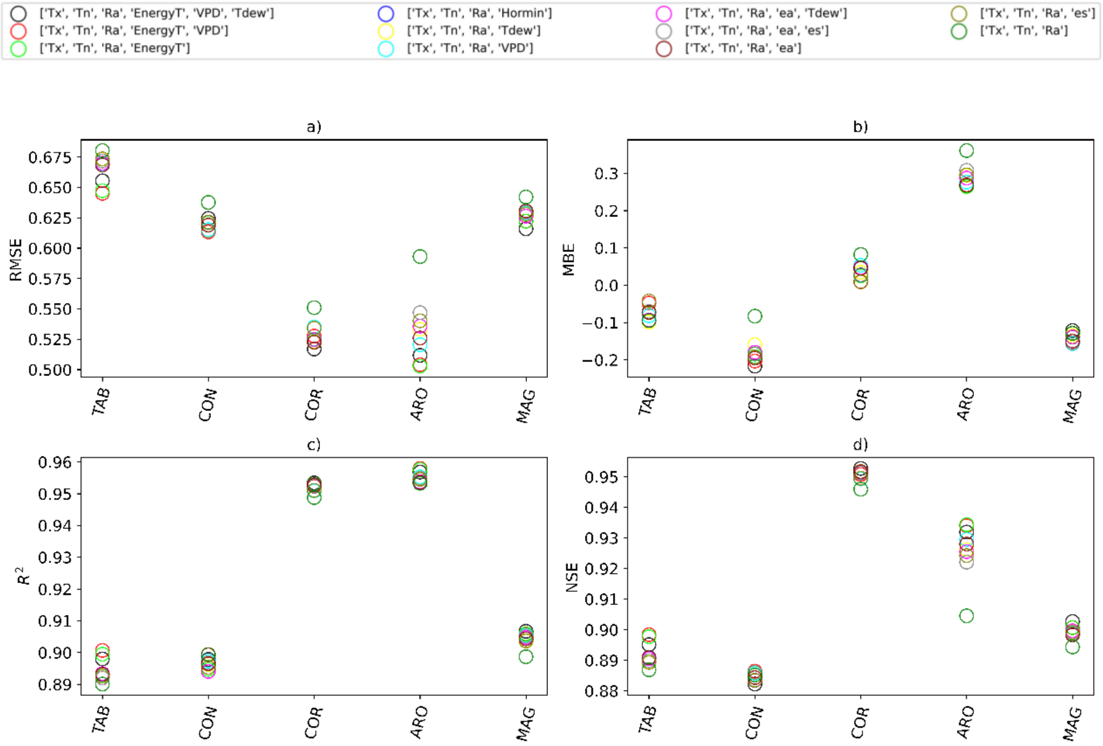


Figure 3. 11. Statistical performances of mean values of RMSE, MBE, R^2 and NSE for all the models evaluated using different configurations (colors) at each location (TAB=Tabernas; CON=Conil; COR=Córdoba; ARO=Aroche; MAG=Málaga). a) RMSE, b) MBE, c)R^2 and d) NSE.

In Table 3. 7, it is shown the best NSE and RMSE (mm/day) values for all the stations in the different configurations. In Córdoba, the best NSE and RMSE values were obtained using configuration 11 (NSE = 0.9560 and RMSE = 0.4982 mm/day) and very close to configurations 8 (NSE=0.9553 and RMSE = 0.5026 mm/day) , 3 (NSE = 0.9551 and RMSE = 0.5038 mm/day) and 9 (NSE = 0.9548 and RMSE = 0.5053 mm/day), in this order. The only configuration that did not outperform configuration 1 (using ELM) in terms of NSE was configuration 10 using ELM. In Málaga, the two best results were given using configurations 5 (RMSE = 0.6070 mm/day), 7 (0.6084 mm/day). and 10 (RMSE = 0.6089 mm/day), having very close values one another; these configurations also obtained the best NSE performances (0.9057, 0.9053 and 0.9052, respectively). On the contrary, configurations 2, 3, 4, 8, 9, and 11 worsen configuration 1 in both NSE and RMSE. In Tabernas, the configurations with Energy_T as inputs (2, 8 and 11) gave the best statistics (RMSE = 0.6272 mm/day and NSE = 0.9034; RMSE = 0.6209 mm/day and NSE = 0.9058; RMSE = 0.6267 mm/day and NSE = 0.9012, respectively). The worst values were obtained by the configuration 10 (RMSE = 0.6610 mm/day and NSE = 0.8933) and 1 (RMSE = 0.6607 mm/day and NSE = 0.8934). In Conil, the best configuration performances were 3, 7, 8, 2, and 10, in this order. The use of Hourmin made the model obtain the most efficient estimation in this location, and the use of Energy_T also outperformed configuration 1. Finally, in Aroche, all the configuration outperformed configuration 1, except for configuration 5 and 10. Configurations 4, 6, 8, 11, and 2 obtained the best RMSE and NSE values. This station had the highest improvements in RMSE from HS and HSMOD, while in Conil, it is obtained the highest NSE outperforming.

Table 3. 7. Best NSE and RMSE (mm/day) values for all the stations (TAB=Tabernas; CON=Conil; COR=Córdoba; ARO=Aroche; MAG=Málaga) in the different input configurations (Table 4). The model in parenthesis is the model that obtained the best NSE and RMSE, respectively.

	Córdoba		**Málaga**		**Tabernas**		**Conil**		**Aroche**	
Conf.	**NSE**	**RMSE (mm/day)**	**NSE**	**RMSE (mm/day)**	**NSE**	**RMSE (mm/day)**	**NSE**	**RMSE (mm/day)**	**NSE**	**RMSE (mm/day)**
1	0,9537	0,5058	0.9036	0.6137	0.8934	0.6607	0.8928	0.5959	0.9297	0.5203
	(ELM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(GRNN)	(GRNN)	(SVM)	(SVM)
2	0.9543	0,5083	0.9015	0.6206	0.9039	0.6272	0.8953	0.5891	0.9386	0.4863
	(SVM)	(SVM)	(ELM)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)
3	0.9551	0.5038	0.9009	0.6223	0.8931	0.6618	0.8966	0.5853	0.9350	0.5003
	(SVM)	(SVM)	(MLP)	(MLP)	(RF)	(RF)	(MLP)	(MLP)	(SVM)	(SVM)
4	0.9543	0.5078	0.9023	0.6180	0.8952	0.6552	0.8940	0.5927	0.9452	0.4591
	(SVM)	(SVM)	(ELM)	(ELM)	(ELM)	(ELM)	(GRNN)	(GRNN)	(MLP)	(MLP)
5	0.9539	0.5102	0.9057	0.6070	0.8924	0.6640	0.8937	0.5935	0.9296	0.5205
	(SVM)	(SVM)	(MLP)	(MLP)	(ELM)	(ELM)	(GRNN)	(GRNN)	(SVM)	(SVM)
6	0.9545	0.5067	0.9033	0.6149	0.8955	0.6543	0.8929	0.5956	0.9444	0.4626
	(SVM)	(SVM)	(ELM)	(ELM)	(MLP)	(MLP)	(GRNN)	(GRNN)	(MLP)	(MLP)
7	0.9540	0.5096	0.9053	0.6084	0.8914	0.6670	0.8958	0.5875	0.9307	0.5168
	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(MLP)	(MLP)	(XGBoost)	(XGBoost)
8	0.9553	0.5026	0.8998	0.6259	0.9058	0.6209	0.8955	0.5883	0.9402	0.4800
	(SVM)	(SVM)	(GRNN)	(GRNN)	(ELM)	(ELM)	(GRNN)	(GRNN)	(ELM)	(ELM)
9	0.9548	0.5053	0.9031	0.6156	0.8937	0.6598	0.8934	0.5943	0.9307	0.5165
	(SVM)	(SVM)	(ELM)	(ELM)	(ELM)	(ELM)	(MLP)	(MLP)	(XGBoost)	(XGBoost)
10	0.9535	0.5122	0.9052	0.6089	0.8933	0.6610	0.8951	0.5894	0.9282	0.5259
	(ELM)	(ELM)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)
11	0.9560	0.4982	0.9012	0.6214	0.9041	0.6267	0.8923	0.5975	0.9394	0.4828
	(SVM)	(SVM)	(SVM)	(SVM)	(ELM)	(ELM)	(GRNN)	(GRNN)	(SVM)	(SVM)

In general, it could be sum up that the use of temperature-based variables outperformed configuration 1; the use of Energy$_T$ is highly recommended due to its efficiency improvements. On the other hand, the use of Hourmin did not have considerable gains in most of the assessed stations, although the results in the less arid stations (Córdoba, Conil, and Aroche) where the aridity index is above 0.45 outperform from the use of T_X, Tn, and R_a (this value varies more in these stations due to a more frequent raining events)

3.3.Seasonal performance behavior

Regarding the performance of the models in different seasons, Figure 3.12 shows the statistical values of RMSE, MBE, $R^{2,}$ and NSE (a, b, c, and d, respectively). In terms of R^2, winter had the worst results for all the studied models, not only on average but also as the highest interquartile range, whereas in autumn the highest precision was obtained. Regarding a ranking of general good performance, the best results were obtained in autumn, followed by spring and summer for both R^2 and NSE values. On the other side, the lowest R^2 value (0.0976) resulted in Málaga with ELM and using configuration 8 (also corresponding with the value NSE= -1.0066, RMSE=0.9246, and MBE=-0.1749). Moreover, MBE did not have any significant variability in the mean values due to seasons, but the interquartile range was more compact in summer for all the models except for MLP and some outliers in Aroche. Finally, concerning RMSE values, the best performance was obtained using the MLP model in Winter, followed by the same approach in Autumn. Regarding all the models and sites studied, in Spring and Summer RMSE values were lower than 0.9 mm.

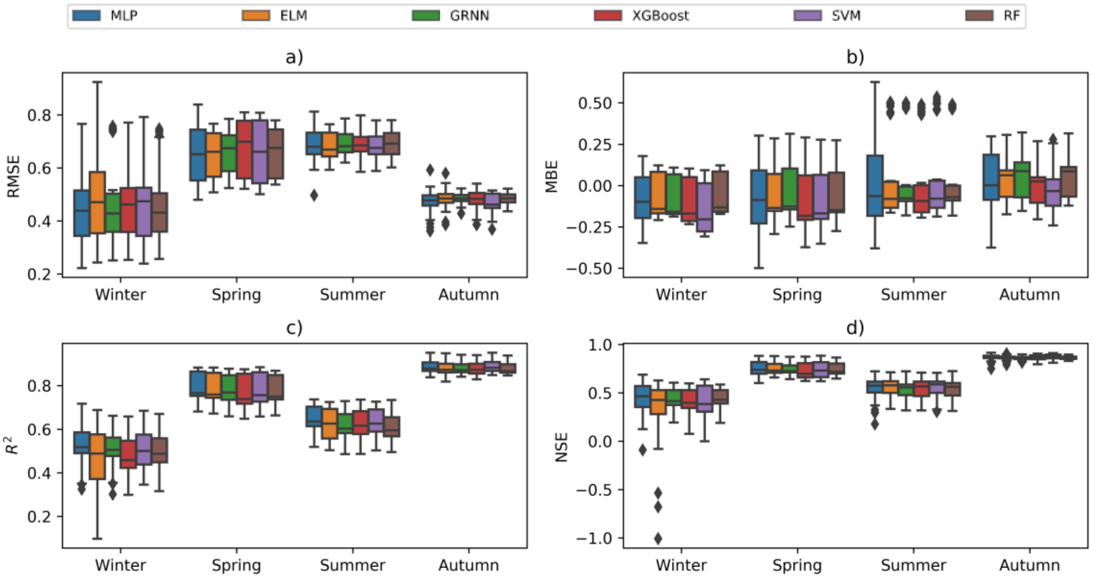


Figure 3. 12. Seasonal statistical performances of RMSE, MBE, R^2, and NSE for all the configurations using different models (colors) in all stations. a) RMSE, b) MBE, c) R^2 and d) NSE.

In Table 3. 8, it is shown the different mean NSE and RMSE values along all models and stations to seasonally assess the performance of the different input configurations. Generally, asit was stated in figure 12, the NSE had better values in Autumn, whereas in winter, the performance was down to the worst. In winter, the best configurations were configuration 6 (NSE=0.4142), 3 (NSE=0.4130) and 9 (NSE=0.4133), respectively. In spring, the best configurations were 2 (NSE=0.7603), 8 (NSE=0.7572) and 11 (NSE=0.7538). In summer, the results are very similar to Spring, where the best configurations are 2 (NSE=0.5675), 3 (NSE=0.5658), 8 (NSE=0.5600), and 11 (NSE=0.5550). Finally, in autumn, there is no significant outperforming among the configurations, is also the most accurate season in all cases.

Table 3. 8. Seasonal mean NSE and RMSE values of all the models and stations in the different input configurations (table 4).

	NSE				**RMSE (mm/day)**			
Conf.	**Winter**	**Spring**	**Summer**	**Autumn**	**Winter**	**Spring**	**Summer**	**Autumn**
1	0.4082	0.7470	0.5498	0.8654	0.4648	0.6606	0.6857	0.4788
2	0.3880	0.7603	0.5675	0.8680	0.4662	0.6443	0.6736	0.4737
3	0.4130	0.7487	0.5658	0.8661	0.4633	0.6576	0.6753	0.4774
4	0.4061	0.7493	0.5385	0.8653	0.4662	0.6583	0.6934	0.4795
5	0.4029	0.7480	0.5401	0.8643	0.4671	0.6595	0.6909	0.4811
6	0.4142	0.7507	0.5466	0.8657	0.4627	0.6563	0.6875	0.4784
7	0.4097	0.7469	0.5349	0.8625	0.4653	0.6611	0.6956	0.4834
8	0.2281	0.7572	0.5600	0.8676	0.5005	0.6483	0.6783	0.4748
9	0.4133	0.7480	0.5425	0.8658	0.4635	0.6591	0.6899	0.4786
10	0.4054	0.7443	0.5403	0.8644	0.4662	0.6639	0.6923	0.4805
11	0.3865	0.7538	0.5550	0.8655	0.4685	0.6514	0.6816	0.4771

The use of configurations with Energy_T and Hourmin obtained significantly outperforms in winter, spring, and summer among all stations and models, so the use of these variables is highly recommended to outperform temperature-based models. It is of interest that Hourmin obtained significant gains in winter when the rains are more frequent in these stations

3.4.Overall discussion

In general, the results obtained by applying the proposed models outperformed the Hargreaves methods and gave slightly better values of NSE and R^2 (NSE=0.82 and R^2=0.83 for the local scenario and daily inputs, and NSE=0.89 and R^2=0.90 for the local scenario and hourly inputs) than those reported by Ferreira and da Cunha [50] for ten weather stations in Brazil (3.93 mm/day of mean ET_0), although slightly worse in terms of mean RMSE values (RMSE=0.53 for the local scenario and daily inputs, and RMSE=0.41 for the local scenario and hourly inputs). However, the RMSE values obtained for most models were better than the results for temperature-based models assessed by Fan et al. [35] in several locations in China (the ET_0 in the studied stations was ranging from 2.5 to 3.4 mm/day in the different stations), with a mean RMSE performance of 0.7069 mm/day with the temperature-based models. What is more, the NSE performance in all the stations with different climate conditions (semi-arid, arid, sub-humid, and humid) of India by Adamala [55] using Wavelet Neural Networks (WNN) are lower than the mean NSE values obtained in the assessed stations of Andalusia. The results of Traore [83] in the Sudano-Sahelian zone gave similar mean R^2

values in the temperature-based models, although the RMSE values were lower in this zone. Besides, the results of RF and GEP obtained by Karimi et al. [84] and Shiri [85] in Iran using temperature-based models were worst in terms of maximum and mean NSE values than those obtained in these 5 stations of Andalusia. Finally, in terms of RMSE and comparing the results with other studies carried out in Spain, the values were lower than those of Landeras et al. [86] with MLP in the Basque Country (Northern Spain) and worse than the results of Martí et al. [69] with calibrated HS in eastern Spain.

In general, the computation cost of ELM and GRNN models were lower than using MLP, XGBoost, and RF; being the ELM approach had one of the best mean performances in all the stations assessed. On the other hand, RF and XGBoost, despite needing the most computational requirements and being slower, on average, it is not reflected that higher computational effort on training in the final results.

The good results obtained when using Energy_T as an input variable in any of the configurations (2, 8, and 11) demonstrates that the information contained in it gives a detailed characterization of the thermic behavior at each location, resulting in more efficient models than using only daily maximum and minimum temperature. Its use in inland stations showed one of the best performances, while its use on coastal stations was not such substantial, which may be due to the effect of the sea (the temperature variability during a day is less affected by meteorological events). On the other hand, Hourmin was also assessed, with no such significant outperforming as Enegy_T had, but improving the modeling from the use of only T_X, Tn, and R_a in the less arid stations (aridity index above 0.45) of Córdoba, Conil, and Aroche (see Table 3. 1). Another arising feature in this work is that the use of temperature-based approximations of e_a, e_s, VPD, and T_{dew}, had improvements over the models using only configuration 1. Although it is usually considered to be learned by the ML model in the ‘training’ process, the introduction of these parameters as input variables may help the model learn, minimizing the possibility of overfitting occurrence.

Therefore, configurations using the new parameters Energy_T and Hourmin can be used in this kind of model to estimate ET_0 more accurately, being ELM model, the one that obtained, in general, the highest performances. This is crucial in stations where the data acquisition is limited, like developing countries or low-cost weather stations that cannot collect the rest of the meteorological variables required by FAO56-PM.

4. Conclusions

Different models using several meteorological temperature-based parameters as input variables have been developed and assessed at five locations in the Andalusia region (Southern Spain), covering different climatic conditions. The main novelty of this work is the use of new meteorological parameters based on intra-daily temperature variables (Energy_T and Hourmin), the evaluation of different configurations (the combination of other climatological parameters such as VPD, e_a, e_s, and T_{dew} as inputs in the different approaches studied), as well as the use of Bayesian optimization to tune hyperparameters for each of them. The results indicated that the use of the novel's temperature-based parameters in these neural network approaches outperformed, in general, the estimations of ET_0 in all the assessed stations, comparing to Hargreaves methods and models using T_x, T_n, and R_a. In terms of RMSE and R^2, the values were reduced from 0.8264 mm/day (HSMOD) to 0.5853 mm/day (MLP) and from 0.8182 (HSMOD) to 0.9039 (MLP), respectively. The improvement in the NSE values were slightly better than in R^2, increasing from 0.7939 (HSMOD) to 0.8967 (MLP). Finally, the best improvement was obtained in the MBE statistic, gaining from -0.2830 mm/day (HSMOD) to -0.0048 mm/day (MLP).

As a rule, there is not a model that firmly outperforms the rest of the models in terms of mean statistics, although MLP was usually the model that gave the best performance in most locations, although its variability was high. Besides, GRNN and ELM were the models with the lowest computational cost, where ELM performed with one of the best NSE and RMSE in most stations and configurations. Additionally, based on the results obtained in this work it can be observed an improvement of the statistical performance by using Energy_T in any of the configurations assessed while using Hourmin had improvements in the most humid stations where the aridity index is above 0.45. Additionally, in terms of seasonal performances, the statistics obtained significantly outperforms in winter, spring, and summer among all stations and models. Therefore, the ELM model with EnergyT and Hourmin can be used to improve limited-climatic models to increase the accuracy of ET_0 estimations.

Further studies can deeply explore these parameters and models for regional scenarios and in locations with different climatic conditions from those studied in this work. Due to the rise of low-cost sensorization, especially for variables such as air temperature, and the possibility of collecting these records with a higher sampling frequency than the conventional daily scale, the approaches proposed in this work may result in greater efficiency in the use of water resources.

5. References

[1] Allen R, Pereira L, Raes D, Fao MS-, Rome undefined, 1998 undefined. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. ScscourtOrg n.d.

[2] Pereira LS, Allen RG, Smith M, Raes D. Crop evapotranspiration estimation with FAO56: Past and future. Agric Water Manag 2015;147:4–20. https://doi.org/10.1016/J.AGWAT.2014.07.031.

[3] Debnath S, Adamala S, Raghuwanshi NS. Sensitivity Analysis of FAO-56 Penman-Monteith Method for Different Agro-ecological Regions of India. Environmental Processes 2015;2:689–704. https://doi.org/10.1007/S40710-015-0107-1/TABLES/4.

[4] Estévez J, Gavilán P, Berengena J. Sensitivity analysis of a Penman–Monteith type equation to estimate reference evapotranspiration in southern Spain. Hydrol Process 2009;23:3342–53. https://doi.org/10.1002/HYP.7439.

[5] Itenfisu D, Elliott RL, Allen RG, Walter IA. Comparison of Reference Evapotranspiration Calculations as Part of the ASCE Standardization Effort. Journal of Irrigation and Drainage Engineering 2003;129:440–8. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:6(440).

[6] Gavilán P, Estévez J, Berengena J. Comparison of Standardized Reference Evapotranspiration Equations in Southern Spain. Journal of Irrigation and Drainage Engineering 2008;134:1–12. https://doi.org/10.1061/(ASCE)0733-9437(2008)134:1(1).

[7] Córdova M, Carrillo-Rojas G, Crespo P, Wilcox B, Célleri R. Evaluation of the Penman-Monteith (FAO 56 PM) Method for Calculating Reference Evapotranspiration Using Limited Data. Https://DoiOrg/101659/MRD-JOURNAL-D-14-00241 2015;35:230–9. https://doi.org/10.1659/MRD-JOURNAL-D-14-0024.1.

[8] Allen RG, Walter IA, Elliott RL, Howell TA, Itenfisu D, Jensen ME, et al. The ASCE Standardized Reference Evapotranspiration Equation. ASCE Standardized Reference Evapotranspiration Equation 2005:1–203. https://doi.org/10.1061/9780784408056.

[9] Estévez J, Padilla FLM, Gavilán P. Evaluation and Regional Calibration of Solar Radiation Prediction Models in Southern Spain. Journal of Irrigation and Drainage Engineering 2012;138:868–79. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000493.

[10] De Paola F, Giugni M. Coupled Spatial Distribution of Rainfall and Temperature in USA. Procedia Environ Sci 2013;19:178–87. https://doi.org/10.1016/J.PROENV.2013.06.020.

[11] WMO. Guide to instruments and methods of observations. Geneva, Switzerland, 2018; Volume 8 2008;I & II:I.8-1 to I.9-1.

[12] Library WO, Eccel E. Estimating air humidity from temperature and precipitation measures for modelling applications. Meteorological Applications 2012;19:118–28. https://doi.org/10.1002/MET.258.

[13] Hargreaves GH, Samani ZA. Reference Crop Evapotranspiration from Temperature. Appl Eng Agric 1985;1:96–9. https://doi.org/10.13031/2013.26773.

[14] Gavilán P, Lorite IJ, Tornero S, Berengena J. Regional calibration of Hargreaves equation for estimating reference ET in a semiarid environment. Agric Water Manag 2006;81:257–81. https://doi.org/10.1016/J.AGWAT.2005.05.001.

[15] Rahimi Khoob A. Comparative study of Hargreaves's and artificial neural network's methodologies in estimating reference evapotranspiration in a semiarid environment. Irrig Sci 2008;26:253–9. https://doi.org/10.1007/S00271-007-0090-Z/FIGURES/5.

[16] Martínez-Cob A, Tejero-Juste M. A wind-based qualitative calibration of the Hargreaves ET0 estimation equation in semiarid regions. Agric Water Manag 2004;64:251–64. https://doi.org/10.1016/S0378-3774(03)00199-9.

[17] Ravazzani G, Corbari C, Morella S, Gianoli P, Mancini M. Modified Hargreaves-Samani Equation for the Assessment of Reference Evapotranspiration in Alpine River Basins. Journal of Irrigation and Drainage Engineering 2012;138:592–9. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000453.

[18] Tang P, Xu B, Gao Z, Li H, Gao X, Wang C. Estimating reference crop evapotranspiration with elevation based on an improved HS model. Hydrology Research 2019;50:187–99. https://doi.org/10.2166/NH.2018.022.

[19] Trajkovic S. Temperature-Based Approaches for Estimating Reference Evapotranspiration. Journal of Irrigation and Drainage Engineering 2005;131:316–23. https://doi.org/10.1061/(ASCE)0733-9437(2005)131:4(316).

[20] Vanderlinden K, Giráldez J V., Van Meirvenne M. Assessing Reference Evapotranspiration by the Hargreaves Method in Southern Spain. Journal of Irrigation and Drainage Engineering 2004;130:184–91. https://doi.org/10.1061/(ASCE)0733-9437(2004)130:3(184).

[21] Todorovic M, Karic B, Pereira LS. Reference evapotranspiration estimate with limited weather data across a range of Mediterranean climates. J Hydrol (Amst) 2013;481:166–76. https://doi.org/10.1016/J.JHYDROL.2012.12.034.

[22] Blaney HF, 1892-. Determining water requirements in irrigated areas from climatological and irrigation data 1952. https://doi.org/10.3/JQUERY-UI.JS.

[23] Thornthwaite. Instructions and tables for computing potential evapotranspiration and the water balance. Publications on Climatology 1957;10:185–310. https://doi.org/10.18960/SEITAI.46.1_73.

[24] Schendel U. Vegetationswasserverbrauch und-wasserbedarf - Google Scholar. J Habilitation 1967;137.

[25] Priestley CHB, Taylor RJ. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters in: Monthly Weather Review Volume 100 Issue 2 (1972). Mon Weather Rev 1972. https://doi.org/10.1175/1520-0493(1972)100%3C0081:OTAOSH%3E2.3.CO;2.

[26] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric For Meteorol 1984;31:159–66. https://doi.org/10.1016/0168-1923(84)90017-0.

[27] Bormann H. Sensitivity analysis of 18 different potential evapotranspiration models to observed climatic change at German climate stations. Clim Change 2011;104:729–53. https://doi.org/10.1007/S10584-010-9869-7/METRICS.

[28] Acheampong PK. Evaluation of potential evapotranspiration methods for Ghana. GeoJournal 1986;12:409–15. https://doi.org/10.1007/BF00262364/METRICS.

[29] Xu C-Y, Singh VP. Evaluation and generalization of radiation-based methods for calculating evaporation n.d. https://doi.org/10.1002/(SICI)1099-1085(20000215)14:2.

[30] McKenney MS, Rosenberg NJ. Sensitivity of some potential evapotranspiration estimation methods to climate change. Agric For Meteorol 1993;64:81–110. https://doi.org/10.1016/0168-1923(93)90095-Y.

[31] Moratiel R, Bravo R, Saa A, Tarquis AM, Almorox J. Estimation of evapotranspiration by the Food and Agricultural Organization of the United Nations (FAO) Penman-Monteith temperature (PMT) and Hargreaves-Samani (HS) models under temporal and spatial criteria - a case study in Duero basin (Spain). Natural Hazards and Earth System Sciences 2020;20:859–75. https://doi.org/10.5194/NHESS-20-859-2020.

[32] Jahanbani H, El-Shafie AH. Application of artificial neural network in estimating monthly time series reference evapotranspiration with minimum and maximum temperatures. Paddy and Water Environment 2011;9:207–20. https://doi.org/10.1007/S10333-010-0219-1/TABLES/2.

[33] Kumar M, Raghuwanshi NS, Singh R, Wallender WW, Pruitt WO. Estimating Evapotranspiration using Artificial Neural Network. Journal of Irrigation and Drainage Engineering 2002;128:224–33. https://doi.org/10.1061/(ASCE)0733-9437(2002)128:4(224).

[34] Martí P, Royuela A, Manzano J, Palau-Salvador G. Generalization of ETo ANN Models through Data Supplanting. Journal of Irrigation and Drainage Engineering 2010;136:161–74. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000152.

[35] Fan J, Yue W, Wu L, Zhang F, Cai H, Wang X, et al. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agric For Meteorol 2018;263:225–41. https://doi.org/10.1016/J.AGRFORMET.2018.08.019.

[36] Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. J Hydrol (Amst) 2019;572:556–70. https://doi.org/10.1016/J.JHYDROL.2019.03.028.

[37] Guo X, Sun X, Ma J. Prediction of daily crop reference evapotranspiration (ET0) values through a least-squares support vector machine model. Hydrology Research 2011;42:268–74. https://doi.org/10.2166/NH.2011.072.

[38] Kisi O. Least squares support vector machine for modeling daily reference evapotranspiration. Irrig Sci 2013;31:611–9. https://doi.org/10.1007/S00271-012-0336-2/TABLES/3.

[39] Shrestha NK, Shukla S. Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. Agric For Meteorol 2015;200:172–84. https://doi.org/10.1016/J.AGRFORMET.2014.09.025.

[40] Tabari H, Kisi O, Ezani A, Hosseinzadeh Talaee P. SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. J Hydrol (Amst) 2012;444–445:78–89. https://doi.org/10.1016/J.JHYDROL.2012.04.007.

[41] Wen X, Si J, He Z, Wu J, Shao H, Yu H. Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration With Limited Climatic Data in Extreme Arid Regions. Water Resources Management 2015;29:3195–209. https://doi.org/10.1007/S11269-015-0990-2/TABLES/5.

[42] Feng Y, Cui N, Gong D, Zhang Q, Zhao L. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. Agric Water Manag 2017;193:163–73. https://doi.org/10.1016/J.AGWAT.2017.08.003.

[43] Kisi O, Kilic Y. An investigation on generalization ability of artificial neural networks and M5 model tree in modeling reference evapotranspiration. Theor Appl Climatol 2016;126:413–25. https://doi.org/10.1007/S00704-015-1582-Z/FIGURES/7.

[44] Pal M, Deswal S. M5 model tree based modelling of reference evapotranspiration. Hydrol Process 2009;23:1437–43. https://doi.org/10.1002/HYP.7266.

[45] Rahimikhoob A. Comparison between M5 model tree and neural networks for estimating reference evapotranspiration in an arid environment. Water Resources Management 2014;28:657–69. https://doi.org/10.1007/S11269-013-0506-X/FIGURES/7.

[46] Abdullah SS, Malek MA, Abdullah NS, Kisi O, Yap KS. Extreme Learning Machines: A new approach for prediction of reference evapotranspiration. J Hydrol (Amst) 2015;527:184–95. https://doi.org/10.1016/J.JHYDROL.2015.04.073.

[47] Feng Y, Peng Y, Cui N, Gong D, Zhang K. Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. Comput Electron Agric 2017;136:71–8. https://doi.org/10.1016/J.COMPAG.2017.01.027.

[48] Gocic M, Petković D, Shamshirband S, Kamsin A. Comparative analysis of reference evapotranspiration equations modelling by extreme learning machine. Comput Electron Agric 2016;127:56–63. https://doi.org/10.1016/J.COMPAG.2016.05.017.

[49] Kim S, Kim HS. Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modeling. J Hydrol (Amst) 2008;351:299–317. https://doi.org/10.1016/J.JHYDROL.2007.12.014.

[50] Ferreira LB, da Cunha FF. New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 2020;234:106113. https://doi.org/10.1016/J.AGWAT.2020.106113.

[51] Shiri J, Kişi Ö, Landeras G, López JJ, Nazemi AH, Stuyt LCPM. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). J Hydrol (Amst) 2012;414–415:302–16. https://doi.org/10.1016/J.JHYDROL.2011.11.004.

[52] Shiri J, Nazemi AH, Sadraddini AA, Landeras G, Kisi O, Fakheri Fard A, et al. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput Electron Agric 2014;108:230–41. https://doi.org/10.1016/J.COMPAG.2014.08.007.

[53] Karimaldini F, Shui LT, Mohamed TA, Abdollahi M, Khalili N. Daily Evapotranspiration Modeling from Limited Weather Data by Using Neuro-Fuzzy Computing Technique. Journal of Irrigation and Drainage Engineering 2011;138:21–34. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000343.

[54] Keshtegar B, Kisi O, Ghohani Arab H, Zounemat-Kermani M. Subset Modeling Basis ANFIS for Prediction of the Reference Evapotranspiration. Water Resources Management 2018;32:1101–16. https://doi.org/10.1007/S11269-017-1857-5/FIGURES/8.

[55] Adamala S. Temperature based generalized wavelet-neural network models to estimate evapotranspiration in India. Information Processing in Agriculture 2018;5:149–55. https://doi.org/10.1016/J.INPA.2017.09.004.

[56] Norman H. Automatic weather stations to boost Ghanaian meteorological capabilities. Meteorol Technol Int 2019.

[57] Nsabagwa M, Byamukama M, Kondela E, Otim JS. Towards a robust and affordable Automatic Weather Station. Dev Eng 2019;4:100040. https://doi.org/10.1016/J.DEVENG.2018.100040.

[58] Strigaro D, Cannata M, Antonovic M. Boosting a Weather Monitoring System in Low Income Economies Using Open and Non-Conventional Systems: Data Quality Analysis. Sensors 2019, Vol 19, Page 1185 2019;19:1185. https://doi.org/10.3390/S19051185.

[59] Mesas-Carrascosa FJ, Verdú Santano D, Meroño JE, Sánchez de la Orden M, García-Ferrer A. Open source hardware to monitor environmental parameters in precision agriculture. Biosyst Eng 2015;137:73–83. https://doi.org/10.1016/J.BIOSYSTEMSENG.2015.07.005.

[60] Estévez J, Bellido-Jiménez JA, Liu X, García-Marín AP. Monthly Precipitation Forecasts Using Wavelet Neural Networks Models in a Semiarid Environment. Water 2020, Vol 12, Page 1909 2020;12:1909. https://doi.org/10.3390/W12071909.

[61] Middleton N, Thomas DSG. World atlas of desertification. World Atlas of Desertification 1992. https://doi.org/10.2307/3060449.

[62] Hossein Kazemi M, Shiri J, Marti P, Majnooni-Heris A. Assessing temporal data partitioning scenarios for estimating reference evapotranspiration with machine learning techniques in arid regions. J Hydrol (Amst) 2020;590:125252. https://doi.org/10.1016/J.JHYDROL.2020.125252.

[63] Shiri J, Marti P, Karimi S, Landeras G. Data splitting strategies for improving data driven models for reference evapotranspiration estimation among similar stations. Comput Electron Agric 2019;162:70–81. https://doi.org/10.1016/J.COMPAG.2019.03.030.

[64] Estévez J, Gavilán P, Giráldez J V. Guidelines on validation procedures for meteorological data from automatic weather stations. J Hydrol (Amst) 2011;402:144–54. https://doi.org/10.1016/J.JHYDROL.2011.02.031.

[65] Martí P, Zarzo M, Vanderlinden K, Girona J. Parametric expressions for the adjusted Hargreaves coefficient in Eastern Spain. J Hydrol (Amst) 2015;529:1713–24. https://doi.org/10.1016/J.JHYDROL.2015.07.054.

[66] Wang Y-M, Traore S, Kerh T. Neural Network Approach for Estimating Reference Evapotranspiration from Limited Climatic Data in Burkina Faso n.d.

[67] Estévez J, García-Marín AP, Morábito JA, Cavagnaro M. Quality assurance procedures for validating meteorological input variables of reference evapotranspiration in mendoza province (Argentina). Agric Water Manag 2016;172:96–109. https://doi.org/10.1016/J.AGWAT.2016.04.019.

[68] Althoff D, dos Santos RA, Bazame HC, da Cunha FF, Filgueiras R. Improvement of Hargreaves–Samani Reference Evapotranspiration Estimates with Local Calibration. Water 2019, Vol 11, Page 2272 2019;11:2272. https://doi.org/10.3390/W11112272.

[69] Martí P, González-Altozano P, López-Urrea R, Mancha LA, Shiri J. Modeling reference evapotranspiration with calculated targets. Assessment and implications. Agric Water Manag 2015;149:81–90. https://doi.org/10.1016/J.AGWAT.2014.10.028.

[70] Hinton G, Srivastava N, Swersky K. Neural Networks for Machine Learning Lecture 6a Overview of mini-¬-batch gradient descent n.d.

[71] Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings 2014.

[72] Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning 2013:1139–47.

[73] Huang G Bin, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. Neurocomputing 2006;70:489–501. https://doi.org/10.1016/J.NEUCOM.2005.12.126.

[74] Specht DF. A General Regression Neural Network. IEEE Trans Neural Netw 1991;2:568–76. https://doi.org/10.1109/72.97934.

[75] Firat M, Gungor M. Generalized Regression Neural Networks and Feed Forward Neural Networks for prediction of scour depth around bridge piers. Advances in Engineering Software 2009;40:731–7. https://doi.org/10.1016/J.ADVENGSOFT.2008.12.001.

[76] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88/METRICS.

[77] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324/METRICS.

[78] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining n.d. https://doi.org/10.1145/2939672.

[79] Močkus J. On Bayesian Methods for Seeking the Extremum. Optimization Techniques IFIP Technical Conference 1975:400–4. https://doi.org/10.1007/978-3-662-38527-2_55.

[80] Borji A, Itti L. Bayesian optimization explains human active search. Adv Neural Inf Process Syst 2013;26.

[81] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.

[82] Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. Adv Neural Inf Process Syst 2012;25.

[83] Traore S, Wang YM, Kerh T. Artificial neural network for modeling reference evapotranspiration complex process in Sudano-Sahelian zone. Agric Water Manag 2010;97:707–14. https://doi.org/10.1016/J.AGWAT.2010.01.002.

[84] Karimi S, Shiri J, Marti P. Supplanting missing climatic inputs in classical and random forest models for estimating reference evapotranspiration in humid coastal areas of Iran. Comput Electron Agric 2020;176:105633. https://doi.org/10.1016/J.COMPAG.2020.105633.

[85] Shiri J. Evaluation of FAO56-PM, empirical, semi-empirical and gene expression programming approaches for estimating daily reference evapotranspiration in hyper-arid regions of Iran. Agric Water Manag 2017;188:101–14. https://doi.org/10.1016/J.AGWAT.2017.04.009.

[86] Landeras G, Ortiz-Barredo A, López JJ. Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain). Agric Water Manag 2008;95:553–65. https://doi.org/10.1016/J.AGWAT.2007.12.011.

Chapter 4

Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions

Juan Antonio Bellido-Jiménez[1], Javier Estévez[1], Amanda P. García-Marín[1]

[1] Projects Engineering Area, Department of Rural Engineering, Civil Constructions and Projects Engineering, University of Córdoba, Spain

Abstract

The measure of solar radiation is costly, as well as its maintenance and calibration needs; therefore, reliable datasets are scarce. In this work, several machine learning models to predict solar radiation have been developed and assessed at nine locations (Southern Spain and North Carolina in the USA), representing different geo-climatic conditions (aridity, sea distance, and elevation). As a novelty, due to the ease of providing air temperature measurements, different new input variables from intra-daily temperature datasets were used. According to the results, all the models highly outperformed self-calibrated empirical methods such as Hargreaves-Samani and Bristow-Campbell, with improvements in RMSE ranging from 7.56% in arid climates to 45.65% in humid. Moreover, regarding mean NSE and R^2 values, several inland locations obtained values above 0.9. In summer, the highest statistics for all sites (more than a 60% improvement in NSE and R^2) were obtained, whereas the worst were given in winter (more than an 18% improvement in NSE and R^2). Besides, when assessing the models in different non-used locations with similar climatic characteristics, the reduction in RMSE was from 0.305 W m^{-2} to 0.252 W m^{-2} in a semiarid coastal climate and from 0.344 W m^{-2} to 0.233 W m^{-2} in dry sub-humid climate, compared to Hargreaves-Samani method. Overall, the MLP obtained the highest performance using the new proposed variables in all locations with medium aridity values, whereas, in the aridest and most humid sites, SVM and RF models were preferred. Therefore, the temperature-based models developed in this work can predict solar radiation more accurately than the current ones. This is crucial in locations with no available datasets or missing/low quality and can be used to optimize the determination of the potential locations for solar power plants' construction.

1. Introduction

Deepening and improving renewable energy sources' knowledge is crucial due to the increase in the world's energy demand and the evident anthropogenic impact on the climate system. In this sense, it is confirmed that climate change is already affecting natural ecosystems, people, and livelihoods worldwide, being the greenhouse gas emissions the highest in history [1,2]. Among a significant number of consequences, it can be stated that due to global warming, the ice is melting at the poles by leaps and bounds [2], there are changes in precipitation patterns [3,4], the sea level will rise around 2.43 m by 2100 [2,5], and the hurricanes will become more frequent and more potent [6].

Due to the vital role of conventional energy sources in the current warming world, governments invest in producing electricity as clean, sustainable, and with less global warming emissions as possible. Supplementarily, recent studies concluded that Guidelines [7]using alternative energy sources like solar energy (using photovoltaic and photovoltaic/thermal panels) highly reduces the energy-environmental impact indices of strawberry [8] and barley production [9]. Consequently, installing renewable energy plants is increasing worldwide, primarily solar plants, due to their easy and affordable installation on a high scale [10] and for self-consumption purposes [11]. Spain connected to the grid with almost 4GW of solar photovoltaic (PV) capacity in 2019 [12]. Africa had more than quadrupled its solar PV capacity from 1.5GW at the end of 2014 to 8 GW at the end of 2019, being Benban the largest solar project in Africa. North and South American countries added 21 GW of new solar PV capacity in 2019 [13]. In this way, a proper determination of optimal solar power plant locations is crucial, being involved in several issues such as climate, geomorphological, spatial, environmental [14,15], and socio-politic factors [16]. Temperature is one of the most critical parameters to be considered due to its impact on solar plants' efficiency [17,18], so a balance between irradiance and air temperature is needed. Furthermore, political decisions also have a powerful influence like the land price and the energy price, among others [16]. Thus, it is crucial to assess solar radiation predictions in different climates in order to estimate the economic and energetical profitability of potential locations.

1.1.About the related work

The evaluation of solar energy availability at specific locations needs an accurate estimation of solar radiation values and the analysis of its spatiotemporal variability [19] for different purposes and applications. However, measuring solar radiation is very expensive in terms of installation and maintenance, and it usually contains a high rate of erroneous and missing data due to the lack of regular maintenance and calibration issues [7,20]. In this sense, due to the high cost involved, there is a low geographical density of weather stations with this equipment [21,22]. Different methods for estimating and forecasting solar radiation have been studied and stated in numerous works worldwide. They can be classified into three main groups: the empirical or physics-based models, models based on meteorological satellite images, and the statistical or Machine Learning (ML) models [23].

The physics-based models study the weather and use different meteorological parameters for this purpose. For example, Trnka et al. [24] assessed 7 empirical methods for estimating daily global radiation in several weather stations with other climatic conditions in Austria and the Czech Republic, reporting Ångström-Prescott as the best of all the tested models. Estévez et al. [22] evaluated different empirical models, such as Hargreaves-Samani (HS), Bristow and Campbell (BC), and the Mahmood-Hubbard equation in 56 weather stations in a semiarid environment. Chen et al. [25] analyzed 294 versions of empirical models (classified into 37 groups according to the input variables requirements) in 3 meteorological stations located in a subtropical monsoon climate region in China. Naserpour et al.[26] assessed 21 sunshine-based empirical models for estimating daily solar radiation in 51 weather stations situated in different climate regions of Iran. Hassan et al. [27] studied the performance of new empirical temperature-based models in 10 other locations around Egypt. Furthermore, imaged-based methods use satellite images and sky cameras to track clouds in order to predict solar irradiance. In these lines, Azhari et al. [28] assessed the use of satellite images in predicting solar radiation in 9 stations in Malaysia's tropical environment. Fu and Chen [29] predicted irradiance at 5 min in advance by using features extracted from all-sky images and evaluated in a coastal site in Taiwan, achieving around 22% of Mean Absolute Error (MAE). Finally, ML methods use historical data to characterize trends and learn patterns, respectively. Liu et al. [30] evaluated Support Vector Machine (SVM) models using a Firefly algorithm (FFA) to tune hyperparameters, which is called SVM-FFA, and compared it with 10 empirical models for estimating daily diffuse solar radiation in different climatic zones of China. Yagli et al. [23] assessed 68 ML models in 7 weather stations at 5 different climate zones in the United States, showing that tree-based methods obtained better nRMSE performances in long-term averages and under all-sky conditions. However, under clear-sky conditions, Multilayer Perceptron (MLP) and SVM were found to be superior. Ghimire et al. [31] used the Moderate Resolution Imaging Spectroradiometer (MODIS) with SVM to estimate monthly solar radiation in three Australian locations. Srivastava and Lessman [32] studied the performance of Long-Short Term Memory (LSTM) models in forecasting a day-ahead global horizontal irradiance in 16 weather stations from Europe and 5 from the US. Amiri et al. [33] introduced a multi-task hybrid Evolutionary Neural Network (ENN) model to simultaneously estimate and forecast (1 h in advance) inclined solar irradiation and compared it with independently trained models. The results of multi-task models were promising due to their accurate results, the more straightforward architecture obtained, and the lower computational cost compared to single models. Guijo-Rubio et al. [34] also studied the performance of ENN for predicting solar radiation using data exclusively from satellite-based measurements (avoiding the use of ground station measurements) in Toledo, Spain. The results were compared with SVM and ELM, although ENN gave the most accurate predictions. Tao et al. [35] proposed the use of the Adaptive Neuro-Fuzzy Inference System (ANFIS) with two metaheuristic optimization algorithms, Salp Swarm Algorithm (SSA) and Grasshopper Optimization Algorithm (GOA), to outperform solar radiation prediction in North Dakota, the USA. These metaheuristic algorithms outperformed those obtained by others, such as Grey Wolf Optimizer (GWO) and Particle Swarm Optimization (PSO).

Pang et al. [36] assessed the use of Recurrent Neural Networks (RNN) in Alabama, improving the RMSE values up to 26% compared to MLP.

The number of works assessing ML using limited climatic data (especially those using temperature-based datasets) is minimal compared to articles using empirical methods. In these terms, Feng et al. [37] evaluated the performance of ANN, RF, and Wavelet Neural Network (WNN) and compared the results with 4 empirical temperature-based models (Hargreaves-Samani, Campbell model, Jahani model, and Fan model) for estimating daily solar radiation in 4 weather stations in an arid region of China. The input data used in these models were maximum and minimum air temperature, extraterrestrial solar radiation, and the Day-Of-Year. Sharifi et al. [38] assessed ANN, Gene Expression Programming (GEP), and Wavelet Regression (WR) to estimate daily solar radiation using daily mean clearness index (K_T), temperature range (DTR), theoretical sunshine duration (N), and extraterrestrial radiation (R_a) in humid and semi-humid weather stations in USA and Iran. Jiménez et al. [39] used a hybrid model between empirical and artificial neural networks using only the air temperature, the air relative humidity, and the atmospheric pressure as input variables in 5 weather stations from the province of Tucumán (Argentina), having good results but with the inconvenience of needing three meteorological sensors. It is worth noting that no work goes deeper using only air temperature variables in order to outperform solar radiation predictions. Instead, all of these previously mentioned approaches need several input variables such as relative humidity, daylight hours, or clearness index, among others, which causes the cost to increase due to the devices involved.

1.2.Research gaps and scientific contribution

This work is motivated by the new vital role of low-cost and non-conventional devices for Automated Weather Stations (AWS) due to the development of the Internet of Things (IoT) with new connectivity technology such as 5G and LoRaWAN [40], as well as the lower price for data storing and the higher hardware specifications of devices like Raspberry and Arduino [41]. The efficiency is not compromised, especially using temperature measurements [42–44], helping to obtain low-cost and worldwide interconnected datasets that could be appropriately used for energy applications, climate change studies, ecology, and agricultural sciences, among others. These motivations justify a comprehensive evaluation of novel approaches addressing the following research gaps: i) Even though the number of works assessing different machine and deep learning models to predict/forecast solar radiation is high, there are very few works using temperature as the only input variable; ii) a trial and error technique is widely used to tune hyperparameters, which is not of high efficiency unlike Bayesian optimization, Particle-Swarm optimization or Genetic algorithms, iii) the use of short meteorological series for calibration/validation procedures may not be enough to characterize the climatic pattern of locations and iv) the assessment of the generated models in new non-trained locations, involving its international application performance. With these research gaps in mind, the purpose of this work is to find a novel approach to outperform the estimation of daily solar radiation estimates in different conditions only using temperature data, as well as the assessment of the trained models in new locations with similar geo-climatic characteristics, providing the following contributions/novelties: i) the use of novel variables based on intra-daily temperature data (like $Energy_T$ and Hourmin) to predict daily solar radiation, ii) an assessment of different ML models (including low-computational cost models like ELM and GRNN) using Bayesian optimization to tune their hyperparameters, providing a more accurate evaluation among them, iii) a temporal assessment in terms of the seasonality of all the models studied and iv) an evaluation of some of the developed models in new sites (non-used datasets for training) with similar geo-climatic characteristics.

The analysis was performed using nine real-world datasets representing different geo-climatic conditions (aridity, sea distance, and elevation), including 18 years of datasets with daily and intra-daily temperature values. The results of this work can serve as an essential reference point in order to obtain more accurate solar radiation estimates: i) where there are no resources to install solar radiation sensors (developing countries or regions with lack of funding), ii) at similar climatic conditions in other regions/countries because the 31% of the Earth´s surface is characterized by arid, semiarid, dry sub-humid climates while the 67% is humid, iii) the same models and configurations, as well as the procedures used, could be easily extended and applied in new sites, and iv) to help determine the best places for future solar plant locations without installing high-cost devices and predicting measurement gaps.

The rest of this work is organized as follows. In Section 2, the theoretical overview of empirical temperature-based and the different assessed ML models are described. The information about locations, details of the datasets used, quality assurance tests, hyperparameter tune, data pre-processing, and evaluation metrics are addressed in Section 3. In Section 4, the results are reported and discussed using four points of view, the model, configuration, seasonality performance, and the application of some models in new non-used locations. Finally, Section 5 describes the conclusions achieved in this work.

2. Theoretical overview

In this section, a brief overview of the models used in this work is provided. Firstly, the conceptual basis of empirical solar radiation estimation methods such as Hargreaves-Samani [45,46] and Bristow-Campbell [15,47] is introduced. Moreover, Multilayer Perceptron (MLP)[36], Extreme Learning Machine (ELM) [48,49], Generalized Regression Neural Network (GRNN) [50], Support Vector Machine (SVM) [51,52], Random Forest (RF) [53–55] and Extreme Gradient Boost (XGBoost) [55] are standard methodologies extensively used in this area and others. Thus, it is essential to define and clarify the theoretical background of these algorithms.

In summary, Hargreaves-Samani and Bristow-Campbell equations are widely used parametric models for estimating solar radiation daily, using maximum and minimum temperature as the unique measurements. MLP, ELM, and GRNN are neural network models based on the interaction of a structured group of neurons. SVM models are different from the neural network because their primary function is to search for a hyperplane to contain/separate all the dataset points. Finally, two tree-based and ensemble models were tested, RF and XGBoost.

2.1.Empirical temperature-based models

Numerous scientists have investigated the development of empirical methods to estimate solar radiation using limited climatic data. Some of the most used models worldwide are based on the difference between the maximum and minimum daily air temperature (T_x and T_n, respectively) and the solar radiation incident outside the Earth's atmosphere (R_a). For example, Prescott [56] modified the initial Angstrom model [57], known as the Angstrom-Prescott model (Eq. 1):

$$\frac{R_s}{R_a} = a + b\frac{n}{N} \qquad \text{Eq 4.1}$$

Hargreaves and Samani (HS) [58] also introduced an empirical formula (Eq. 2) for estimating solar radiation using R_a, T_x, T_n, and K_T as a calibration coefficient, being recommended the use of 0.16 and 0.19 in K_T for inland and coastal locations, respectively [59].

$$R_s = K_T\, R_a \sqrt{T_x - T_n} \qquad \text{Eq. 4.2}$$

Another important contribution to temperature-based modeling is the BC approach reported by Bristow and Campbell [60], where R_s is also related to R_a and a function of T_x and T_n (Eq. 3), while a, b, and c are empirical coefficients.

$$R_s = aR_a\left(1 - exp(-b(T_x - T_n)^c)\right) \qquad \text{Eq. 4.3}$$

In this work, a self-calibration procedure was carried out for Hargreaves-Samani, and Bristow-Campbell approaches. Due to the wide use of these equations (México [61], India [62], or China [63]), both models' performance was assessed to be compared to the new machine learning approaches.

2.2.Machine Learning models

All the ML models were developed using python libraries (Keras, Scikit-learn, Scikit-optimize, Tensorflow, XGBoost, Neupy, and Hpelm) under a server workstation with the following features: 2x Intel® Xeon® CPU E5-2650 v3 @ 2.30GHz and 128 GB of RAM. The models' algorithms are detailed below:

2.2.1. Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) model is one of the most common predictors, especially in hydrology and meteorology. For example, gap-filling for river stage data [41], rainfall predictions and forecasts [64–66], temperature [67], and reference evapotranspiration [68,69], among others. The algorithm is based on the biological nervous system, composed of a large number of fully connected neurons, and structured in three kinds of layers: the input, the output, and the hidden layer (see Figure 4. 1). The input and output layers correspond to the inputs and outputs of the model, respectively. For example, in this work, the maximum and minimum temperatures, and the extraterrestrial solar radiation, among other temperature-based variables, were inputs, whereas the solar radiation estimates were the outputs. Moreover, the hidden layer is where the 'neurons' (also known as nodes) are located.

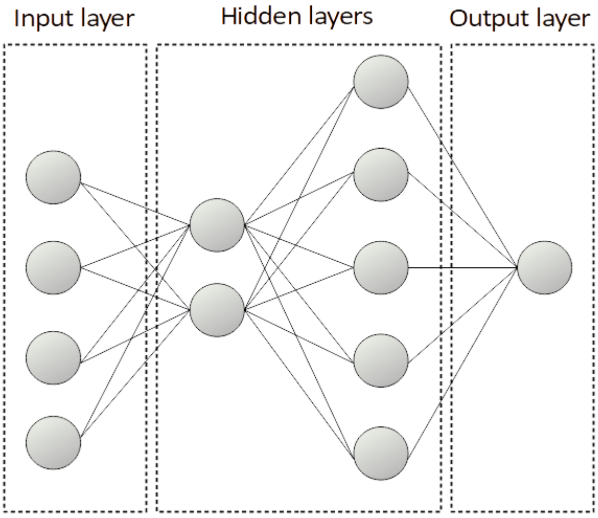


Figure 4. 1. Multilayer Perceptron structure

A single neuron can be mathematically expressed as Eq. 4

$$y = f\left(\sum_{i=1}^{L} w_i x_i + b\right) \qquad \text{Eq. 4.4}$$

During the training process, every neuron's weight and bias in the hidden layer is automatically updated every epoch until the error was minimized, which is why it is called hidden. For further details, the work carried out by Ferreira et al. [68] can be read.

2.2.2. Extreme Learning Machine (ELM)

Huang et al. [70] introduced a new predictor as a single hidden layer feedforward neural network (SLFNN) with fewer computational requirements than MLP. This model's main feature is that the weights and biases are not updated during a training process. On the contrary, the input layer values are randomly generated, and the output values are analytically determined. This behavior makes ELM have a low computational cost because no training iteration is required. Besides, ELM obtains satisfactory results for hydrometeorology [71,72] and energy areas [73,74].

2.2.3. Generalized Regression Neural Network (GRNN)

Generalized Regression Neural Network (GRNN) models are based on radial basis function networks (RBFN) and were first proposed by Specht [74]. Its structure comprises four layers (the input, the pattern, the summation, and the output layer); see Figure 4. 2. The input and output layers represent the inputs and outputs of the model, respectively. The pattern layer contains as many neurons as data in the training dataset; it calculates the distance between the training and the testing data, passing through a radial-based function. The summation layer has two neurons: the S-node and the D-node. The S-node computes the summation of the weighted outputs in the pattern layer, whereas the D-node computes the summation of the unweighted values. Finally, the output can be expressed as the ratio between the S-node output and D-node output. This process is mathematically described as follows (equations 5, 6, and 7) [75]:

$$S = \sum_{i}^{M} w_i \, exp\left(\frac{-(X - X_i')^T (X - X_i')}{2\sigma^2}\right) \qquad \text{Eq. 4.5}$$

$$D = \sum_{i}^{M} exp\left(\frac{-(X - X_i')^T (X - X_i')}{2\sigma^2}\right) \qquad \text{Eq. 4.6}$$

$$Y(X) = \frac{S}{D} \qquad \text{Eq. 4.7}$$

One of this model's unique features is that it does not require an iterated training procedure, so the model is high-speed and gives satisfactory results [71,75].

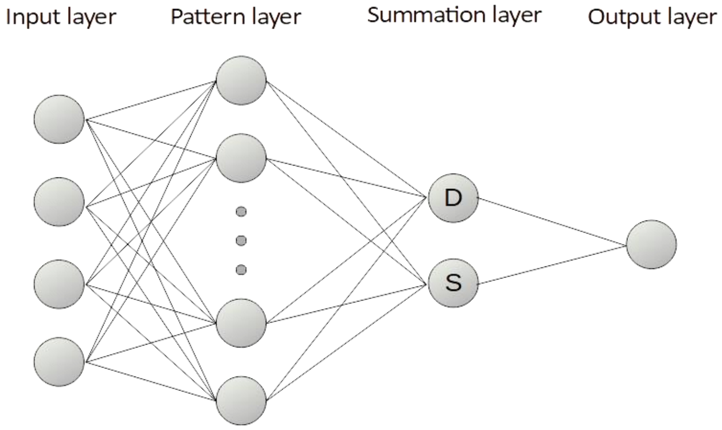
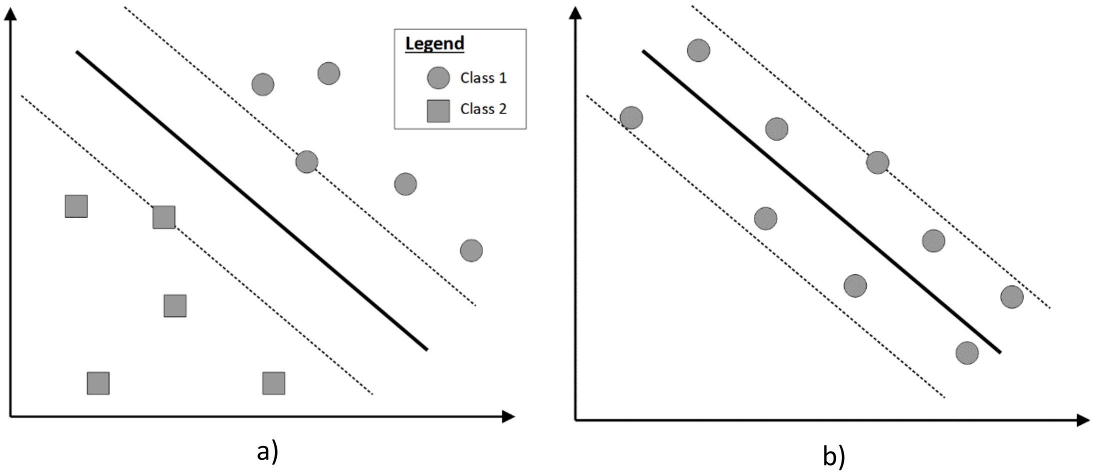


Figure 4. 2. Generalized Regression Neural Network architecture

2.2.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) models are based on a different principle than neural networks like MLP, ELM, and GRNN. In this case, the objective is to search a hyperplane that classifies or contains (for classification or regression tasks, respectively) the data points for an N-dimensional space (where N represents the number of features/inputs of the model). The training process is required for maximizing the margins of the defined hyperplane (see Figure 4. 3).



a) b)

Figure 4. 3. Example of SVM for a) classification and b) regression

The use of SVM in solar radiation predictions has been widely used with promising results in different climates and countries like China [30], the United States [23], Australia [31,76], and Iran [77]. For further details, the following work can be reviewed: [78].

2.2.5. Random Forest (RF)

An ensemble learning algorithm is a kind of ML model where the outputs of a group of predictors are evaluated to obtain a better final performance. It could be seen as a combination of weak learners getting a stronger one. For example, a group of Decision Tree models can be trained on a different random subset of the training dataset, and the final output must consider the predictions of all individual trees. Such an ensemble model is called Random Forest, and it was first proposed by Breiman [79].

RF introduces an extra randomness characteristic when the number of trees starts to grow, in which the node of the tree splits with the best feature among a random subset of them. Therefore, greater tree diversity is obtained. For further details, the work Ibrahim and Khatib [54] can be consulted.

2.2.6. Extreme Gradient Boosting (XGBoost)

Like Random Forest, the Gradient Boosting Machines model (GBM) is classified as an ensemble model, although the main idea in the boosting technique is training to reduce the previous one's error sequentially. In this term, Chen et al. [80] introduced XGBoost as a model to reduce the computational cost for ensemble models and prevent them from overfitting, thanks to executing parallel calculations.

3. Materials and methods

3.1.Source of data

This work was carried out in Andalusia, located in the south of the Iberian Peninsula and southwest of Europe, and in Asheville city from North Carolina, USA. Andalusia is an arid-semiarid region whose main features are: an elevation above mean sea level from 26 to 822 m, the meridians range from 1º to 7º W, the parallels from 37º to 39º N, and a total area of 87,268 m^2 while Asheville is located at 35.49º N, 82.61º W, with an elevation of 650 m. All the sites studied represent 98% of the terrestrial surface. It is important to note that the Andalusia region (Southern Spain) is a very heterogeneous climate area: from Continental climate to Mediterranean (sub-desert, tropical, oceanic, and mountainous).

The datasets used for this work belong to the Agroclimatic Information Network of Andalusia (RIAA) and the National Centers for Environmental Information (NCEI), respectively, and they can be downloaded at https://www.juntadeandalucia.es/agriculturaypesca/i fapa/riaweb/web/ and ftp://ftp.ncei.noaa.gov/pub/data/uscrn/products/. Nine locations were chosen to represent different conditions, according to the United Nations Environment Programme (UNEP) aridity index [81] and corresponding to more than 30% of the Earth's surface to arid, semiarid, and dry sub-humid climate and a 67% to humid [81]. Table 4. 1 shows the different geo-climatic characteristics of the selected Automated Weather Stations (AWS) and the dataset's available period. Figure 4. 4 displays the geographical distribution of all the locations assessed in this work. Table 4. 2 reports the technical specifications of all the sensors used in all the AWS from RIAA measuring temperature and relative humidity of the air, solar radiation, wind speed, wind direction, and precipitation in both intra-daily and daily timescale. Besides, the information about NCEI measuring devices can be found at https://www.ncdc.noaa.gov/crn/instruments.html. Finally, Figure 4. 5 shows the monthly variation of mean daily temperature, wind speed, relative humidity, and solar radiation of all the stations.

Figure 4. 4. Spatial distribution of the weather stations used in this work: a) Almuñécar, Bélmez, Cabra, Córdoba, Écija, El Campillo, Huércal-Overa, Jimena de la Frontera, Loja, Málaga, Mancha Real and Tabernas (Spain) and b) Asheville (USA).

Table 4. 1 Summary of automated weather station sites used in this paper (Lat: Latitude; Lon: Longitude; Altitude: elevation above mean sea level and Aridity Index as the relation Precipitation/ET$_0$ as the annual mean of the whole dataset period).

Station	Aridity index	Altitude (m)	Latitude (ºN)	Longitude (ºW)	Mean annual precipitation (mm)	Time period
Trained/tested stations						
Tabernas (ALM04)	0.1786 (arid)	502	37.091	2.302	243.66	From February 2000 to July 2018
Jimena de la Frontera (CAD07)	0,5975 (dry sub-humid)	50	36.413	5.383	726.94	From November 2000 to July 2018
Córdoba (COR06)	0.4616 (semiarid)	94	37.856	4.802	614.81	From October 2000 to July 2018
Loja (GRA03)	0.3162 (semiarid)	463	37.169	4.138	433.07	From August 2000 to July 2018
El Campillo (HUE08)	0.5497 (dry sub-humid)	383	37.660	6.599	718.67	From November 2000 to July 2018
Mancha Real (JAE07)	0.2808 (semiarid)	407	37.916	3.596	395.38	From August 2000 to July 2018
Málaga (MAG01)	0.3666 (semi-arid)	55	36.756	4.537	457.38	From October 2000 to July 2018
Écija (SEV09)	0.3615 (semi-arid)	109	37.592	5.176	482.17	From November 2000 to July 2018
Asheville (ASH08)	1.1494 (humid)	650	35.490	82.610	1185.67	From November 2000 to December 2018
New sites used to further assess developed models						
Bélmez	0.3681	503	38.254	5.209	496.51	From January 2014 to

(COR01)	(semi-arid)					July 2018
Almuñécar (GRA11)	0.4443 (semi-arid)	29	36,751	3,678	484.50	From January 2014 to July 2018
Cabra (COR101)	0.5641 (dry sub-humid)	543	37.498	4.431	619.76	From January 2014 to July 2018
Huércal-Overa (ALM07)	0.2284 (arid)	303	37.412	1.884	279.39	January 2014 to July 2018

Table 4. 2. Specifications of the RIAA sensors

Sensor device	Measured parameter	Accuracy	Range
Humicap 18'	Relative humidity	2% (from 0 to 90%)	From 0.8 to 100%
		3% (from 90 to 100%)	
Skye SP1110	Solar radiation	±5%	From 350 to 1100 nm
ARG100	Precipitation	0.2 mm/tip	
PT1000	Temperature	±0.3 ºC	From -39.2 to 60 ºC
Young 05103	Wind direction	±3%	From 0 to 360º
	Wind speed	±0.3 m/s	From 0 to 60 m/s

In this work, only measured temperature-based variables were used as inputs, such as the maximum and minimum daily air temperature (T_x and T_n, respectively), the maximum and minimum daily air temperature from the previous day ($T_{x\ prev}$ and $T_{n\ prev}$, respectively), the minimum air temperature of the following early morning ($T_{n\ next}$), the difference between T_x and T_n ($delta_T$), the time at T_x occurs ($Hourmin_{Tx}$), the time at T_n occurs ($Hourmin_{Tn}$), the integral of the half-hourly temperature values during each day ($Energy_T$) [82] and the value of $Energy_T$ of the previous day ($Energy_{T\ prev}$), among others. Finally, the daily extraterrestrial solar radiation (R_a) was also calculated for each location and used as an input variable for all the models assessed.

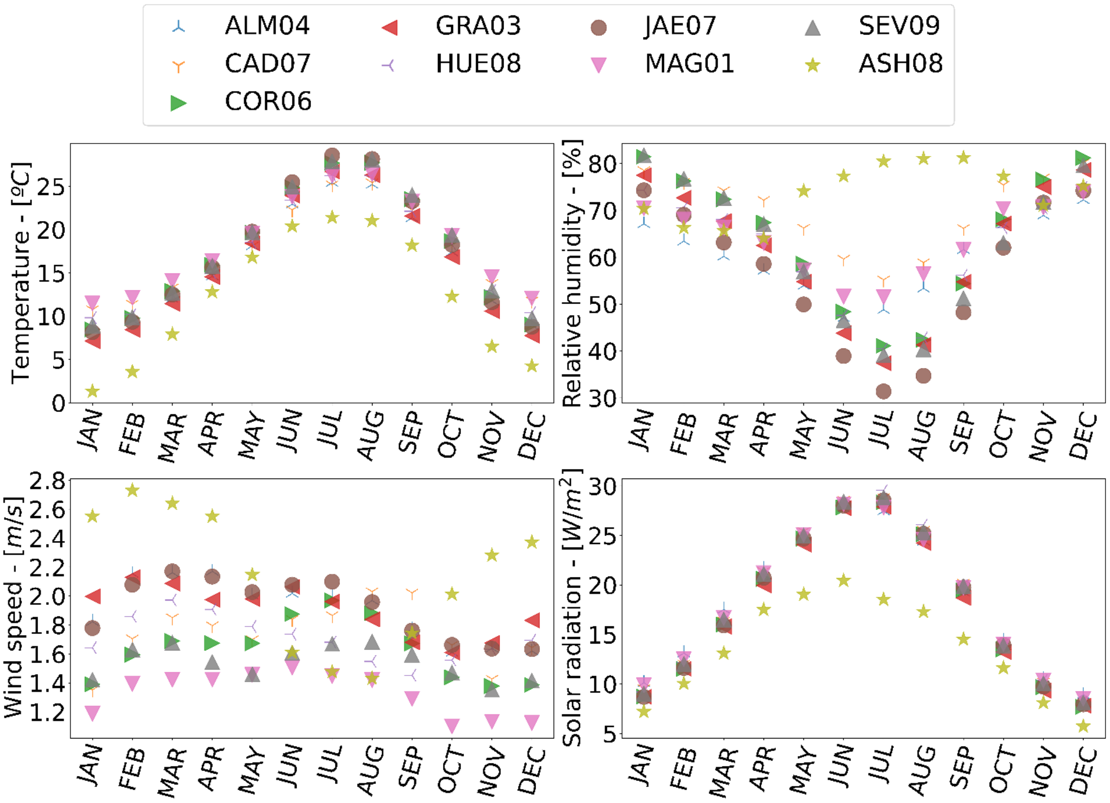


Figure 4. 5. Monthly values of solar radiation, wind speed, relative humidity, and temperature for the weather stations assessed in this study (ASH08 = Asheville, ALM04 = Tabernas, CAD07 = Jimena de la Frontera, COR06 = Córdoba, GRA03 = Loja, HUE08 = El Campillo, JAE07 = Mancha Real, MAG01 = Málaga, SEV09 = Écija)

3.2.Quality assurance procedures

The application of validation procedures is an essential prerequisite to guarantee reliable results when using raw meteorological datasets for different purposes [7]. In this work, the guidelines proposed by Estévez et al. [7,83] were followed by applying a set of quality control procedures (persistence test, internal consistency, range test, and step test), as well as the spatial consistency test proposed by Estévez et al. [84].

3.3.Bayesian optimization

One of the most critical aspects of machine learning is finding the best hyperparameters (architecture and configuration parameters) to obtain the most accurate estimations. There is no general rule to find them, making scientists use a trial and error technique [85]. A more accurate approach uses automatic methods such as Grid or random search [23,86] because all the sets are tested automatically. However, these methodologies are blind to past evaluation results, spending a significant time testing the wrong parameters. In these terms, Bayesian optimization is an efficient solution that addresses this problem. It was first introduced by Mockus [87] as an algorithm that uses the Bayes theorem to select the next set of hyperparameters to test, based on past results.

In Python, the scikit-optimize library has been used to carry out this algorithm, where the hyperparameter hyperspace (the range of the different hyperparameters), as well as the acquisition function (the function that allows the algorithm to determine the next point to evaluate), were defined following the works of Bellido et al. [88,89]. For further details, [90,91] can be revised.

Furthermore, the following step was splitting the whole dataset into two groups: one for the training process (from 2000 to 2014) and the second one for testing (from 2014 to 2018) [66,89]. All the calculations required to tune the different hyperparameters (using Bayesian optimization) were carried out on the training dataset, where 20% of it was used for validation tasks (assessing the performance of different hyperparameter sets). Finally, once optimized, the model was re-trained using the entire training dataset and assessed using the testing dataset. It is essential to highlight that the testing dataset cannot be used to tune hyperparameters because an unwanted bias would be introduced to the final results.

3.4.Data standardization

A common requirement among machine learning models is the standardization of the input dataset, being RF and XGBoost as the only exceptions to this requirement (enhancing the efficiency of models' predictions). Eq. 8 is the mathematical formula followed for this purpose:

$$x^{*} = \frac{x - \bar{\mathrm{x}}}{\sigma_x} \qquad \text{Eq. 4.8}$$

3.5.Statistical analysis

All the models and configurations were evaluated using the following statistics: Mean Bias Error (MBE), Root Mean Squared Error (RMSE), relative Root Mean Squared Error (RRMSE), the Nash-Sutcliffe model efficiency coefficient (NSE), and the coefficient of determination (R^2). All of them are defined in the equations below (Equations. 4.9, 4.10, 4.11, 4.12, and 4.13, respectively).

$$MBE = \frac{1}{m}\sum_{i=1}^{m} meas_i - pred_i \qquad \text{Eq. 4.9}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2} \qquad \text{Eq. 4.10}$$

$$RRMSE = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2}}{\overline{meas}} \qquad \text{Eq. 4.11}$$

$$NSE = 1 - \frac{\sum_{i=1}^{m}(meas_i - pred_i)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2} \qquad \text{Eq. 4.12}$$

$$R^2 = \frac{\left(\sum_{i=1}^{m}(meas_i - \overline{meas})(pred_i - \overline{pred})\right)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2 \sum_{i=1}^{m}(pred_i - \overline{pred})^2} \qquad \text{Eq. 4.13}$$

To get a ranking of all the models assessed related to all the statistics parameters detailed above, a general performance indicator (GPI) was also calculated [92], following Eq. 4.14:

$$GPI_i = \sum_{j=1}^{5} \alpha_j \left(\overline{\mathrm{stats}^{i}} - stats^{ij}\right) \qquad \text{Eq. 4.13}$$

4. Results and Discussion

In order to select the most efficient configurations of different combinations from all the input variables mentioned in section 3.1, a previous ranking was carried out using GPI values for all the stations and using the ELM approach, due to its very satisfactory performance in related works [45]. These numerous configurations always included T_x, T_n, and R_a. Table 4. 3 details the best 14 selected configurations in terms of GPI. These configurations were used for all the models developed in this work. Additionally, the best model from each location was uploaded to an open-access repository in GitHub (https://github.com/Smarity/solar-prediction-only-temp.git), in order to other researchers use them.

Table 4. 3. Input configurations of all the models, where T_x is the maximum daily temperature, T_n is the minimum daily temperature, R_a is the extraterrestrial solar radiation, $Energy_T$ is the integral of the daily thermal signal, $Hourmin_{Tx}$ is the difference between the time when sunset and T_x occurs, and $Hourmin_{Tn}$ is the difference between the time when sunrise and T_n occurs

Conf.	Parameters	Number of inputs
1	T_x, T_n, R_a, and $Hourmin_{Tx}$	4
2	T_x, T_n, R_a, $Energy_T$, and $Hourmin_{Tx}$	5
3	T_x, T_n, R_a, $delta_T$, $Energy_T$, and $Hourmin_{Tx}$	6
4	T_x, T_n, R_a, $Energy_T$, $Hourmin_{Tx}$, and $T_{n\ prev}$	6
5	T_x, T_n, R_a, $Energy_T$, $Hourmin_{Tx}$, and $T_{x\ prev}$	6
6	T_x, T_n, R_a, $Energy_T$, $T_{x\ prev}$, and $T_{n\ next}$	6
7	T_x, T_n, R_a, $Energy_T$, $Hourmin_{Tx}$, $T_{x\ prev}$, and $T_{n\ next}$	7
8	T_x, T_n, R_a, $delta_T$, $Energy_T$, $Hourmin_{Tx}$, and $T_{n\ prev}$	7
9	T_x, T_n, R_a, $delta_T$, $Energy_T$, $Hourmin_{Tx}$, and $T_{x\ prev}$	7
10	T_x, T_n, R_a, $delta_T$, $Energy_T$, $Hourmin_{Tx}$, and $Hourmin_{Tx\ prev}$	7
11	T_x, T_n, R_a, $delta_T$, $Energy_T$, $T_{n\ prev}$, and $T_{n\ next}$	7
12	T_x, T_n, R_a, $Energy_T$, $Energy_T$ prev, $Hourmin_{Tx}$, and $T_{x\ prev}$	7
13	T_x, T_n, R_a, $delta_T$, $Energy_T$, $Hourmin_{Tx}$, $T_{x\ prev}$, and $T_{n\ next}$	8
14	T_x, T_n, R_a, $delta_T$, $Energy_T$, $Energy_{T\ prev}$, $Hourmin_{Tx}$, and $T_{x\ prev}$	8

4.1.Models performance in different locations

Table 4. 4 summarizes the maximum, mean, and minimum values of the statistics MBE, RMSE, NSE, $R^{2,}$ and the mean GPI for the different models and locations. In general, the results obtained by empirical methods (BC and HS) were outperformed by

all ML models for all the statistics and locations. GRNN obtained the worst performances regarding RMSE, NSE, and R^2, contrary to MLP, ELM, and SVM, being the best in most cases. In terms of MBE values, RF and MLP were, on average, the most accurate models.

In Tabernas station (ALM04), which is the aridest location (aridity index=0.178) and with an elevation above mean sea level of 502 m, SVM obtained the best GPI performance (GPI=0.197) followed by MLP (GPI=0.177).

The mean NSE, $R^{2,}$ and RMSE values obtained by SVM and MLP were: NSE=0.867 and 0.865, R^2=0.869 and 0.868, RMSE=0.243 Wm^{-2}, and 0.246 Wm^{-2}, respectively. Besides, the lowest mean MBE value (0.001 Wm^{-2}) was obtained by MLP. On the contrary, in the most humid station from Andalusia (aridity index = 0.5975), Jimena de la Frontera (CAD07), located at 50 m a.m.s.l., RF obtained the highest GPI value (0.312) and almost doubling the result of ELM (GPI=0.187). The best statistics were NSE=0.855 (MLP and RF), R^2=0.871 (SVM), RMSE=0.271 Wm^{-2} (MLP), and MBE=0.052 Wm^{-2} (XGBoost). In Córdoba (COR06), an inland location with an aridity index of 0.461 and an elevation of 94 m, the results were one of the best among all the locations. The NSE and R^2 statistics ranged from NSE=0.879 (XGBoost) and R^2=0.881 (XGBoost) to NSE=0.924 (SVM) and R^2=0.925 (SVM). On average, the best RMSE (0.206 Wm^{-2}) and MBE (0.002 MJ m^{-2}) values were obtained by SVM and XGBoost, respectively, while the best model in terms of GPI was MLP (GPI=0.431). However, in Loja station (GRA03), with an aridity index of 0.316 and 463 m a.m.s.l., the MBE values were, on average, the worst compared to the rest. The best GPI (0.677) was obtained by MLP and the RMSE, NSE, and R^2 values ranged from RMSE=0.243 Wm^{-2} (MLP), NSE=0.855 (MLP), R^2=0.908 (SVM) to RMSE=0.339 Wm^{-2} (ELM), NSE=0.718 (ELM) and R^2=0.824 (ELM). In El Campillo (HUE08), a humid location (aridity index = 0.549) with an elevation of 383 m, MLP was the best model in terms of GPI (0.711). Moreover, this model obtained the best NSE (0.921), R^2 (0.927), RMSE (0.202 Wm^{-2}), and MBE (0.030 Wm^{-2}) values, as well as on average. Similarly, in Mancha Real (JAE07), with an aridity index of 0.280 and an elevation of 407 m, the best GPI was obtained by MLP (GPI=0.644). The statistics obtained by the ML models ranged from NSE=0.888 (GRNN), R^2=0.891 (GRNN), RMSE=0.199 Wm^{-2} (MLP), and MBE=0.000 Wm^{-2} (MLP) to NSE=0.924 (MLP), R^2=0.925 (MLP), RMSE=0.242 Wm^{-2} (GRNN) and MBE=0.042 Wm^{-2} (MLP). In the coastal location of Málaga (MAG01), with an aridity index of 0.366, the best GPI value was also obtained by MLP (GPI=0.480). In addition, the best statistic mean values were obtained by MLP (NSE=0.850, R^2=0.857, RMSE=0.262 Wm^{-2}) and by RF (MBE=0.016 Wm^{-2}). In Écija station (SEV09), with an aridity index of 0.3615 and an altitude of 109 m, MLP and ELM obtained the best GPI performances (GPI=0.367 and GPI=0.291, respectively). The statistics obtained by the ML models evaluated ranged from NSE=0.868 (GRNN), R^2=0.894 (GRNN), RMSE=0.193 Wm^{-2} (MLP), and MBE=0.032 Wm^{-2} (MLP) to NSE=0.924 (MLP), R^2=0.929 (SVM), RMSE=0.255 Wm^{-2} (GRNN) and MBE=0.130 Wm^{-2} (MLP). Finally, in Asheville (ASH08), the most humid site studied in this work, the most accurate performances in terms of RMSE and NSE were carried out by XGBoost (RMSE=0.187 Wm^{-2} and NSE=0.897), followed very closely to RF (RMSE=0.194 Wm^{-2} and NSE=0.890). This site obtained the biggest improvement compared to BC (RMSE=0.300 Wm^{-2}, NSE=0.734) and HS (RMSE=0.344 Wm^{-2} and NSE=0.654), as well as the lowest RMSE value. Regarding MBE, all models and configurations carried out very satisfactory results ranging from MBE=0.001 Wm^{-2} using ELM to 0.026 Wm^{-2} using SVM.

Analyzing the different stations, Córdoba, Écija, Mancha Real, El Campillo, and Asheville obtained the best results on average, in RMSE and NSE values among all models, with RMSE<0.25 Wm^{-2} and NSE>0.9 (with the exception of Asheville, with a mean NSE=0.821). On the contrary, Tabernas, Jimena de la Frontera, Loja, and Málaga obtained worse mean statistics, with RMSE>0.26 Wm^{-2} and 0.8<NSE<0.85. Generally, all ML models outperformed BC and HS in all stations with most of their configurations, obtaining the highest RMSE improvement percentages (as well as the minimum RMSE value) in the most humid site (Asheville) using XGBoost: 37.66 % and 45.64% regarding BC and HS, respectively. Instead, the lowest improvement in this statistic was obtained using ELM in the aridest station (Tabernas): 7.565% and 12.198% regarding BC and HS, respectively. If a more profound look is taken at this aspect, the results show a certain relationship between the best improvements obtained by ML and the aridity index. The higher the aridity index (humid and dry-subhumid climates), the better the results obtained by ML compared to empirical local-calibrated methods. Finally, another essential aspect to highlight is that MLP, SVM, and ELM performed better in arid and semiarid environments, whereas tree-based models like RF and XGBoost were more accurate in sub-humid and humid sites.

Table 4. 4. Maximum (max), arithmetic mean (mean), and minimum (min) values for MBE, RMSE, NSE, and R^2, for all the locations and models (BC = Bristow and Campbell, HS = Hargreaves and Samani, ELM =Extreme Learning Machine, GRNN = Generalised Regression Neural Networks, MLP = Multilayer Perceptron, RF = Random Forest, SVM = Support Vector Machine and XGBoost = Extreme Gradient Boosting) evaluated in this work. For MBE, the minimum and maximum values are in terms of absolute values. The GPI values for each location and model were also reported, and the best mean values of all the statistics are in bold.

		MBE [W/m^2]			**RMSE** [W/m^2]			**NSE**			**R^2**			
Station	**Model**	**max / mean / min**			**max / mean / min**			**max / mean / min**			**max / mean / min**			**GPI**
Tabernas (ALM04)	BC		-0.011			0.256			0.853			0.853		-0.938
	HS		-0.010			0.270			0.837			0.837		-2.423
	ELM	-0.011	-0.006	-0.002	0.251	0.246	0.237	0.874	0.864	0.859	0.874	0.865	0.860	0.109
	GRNN	0.015	0.007	0.002	0.274	0.263	0.256	0.853	0.845	0.832	0.853	0.846	0.834	-1.796
	MLP	-0.061	**0.001**	0.003	0.254	0.246	0.239	0.872	0.865	0.855	0.874	0.868	0.858	0.177
	RF	-0.014	-0.011	-0.008	0.255	0.248	0.242	0.869	0.863	0.855	0.869	0.863	0.855	0.015
	SVM	0.030	0.023	0.015	0.248	**0.243**	0.239	0.872	**0.867**	0.862	0.875	**0.869**	0.865	**0.197**
	XGBoost	-0.011	-0.007	-0.003	0.267	0.250	0.239	0.872	0.861	0.840	0.872	0.861	0.842	-0.232
Jimena de la Frontera (CAD07)	BC		0.093			0.331			0.784			0.802		-3.060
	HS		0.088			0.342			0.770			0.788		-3.585
	ELM	0.083	0.073	0.068	0.288	0.281	0.275	0.851	0.845	0.836	0.865	0.860	0.853	0.187
	GRNN	0.093	0.088	0.081	0.298	0.290	0.282	0.843	0.834	0.825	0.861	0.853	0.844	-0.572
	MLP	0.103	0.075	0.055	0.294	0.281	0.271	0.855	0.844	0.829	0.868	0.861	0.852	0.127
	RF	0.074	**0.071**	0.067	0.285	**0.278**	0.272	0.855	**0.847**	0.840	0.867	0.859	0.852	**0.312**
	SVM	0.104	0.095	0.078	0.298	0.286	0.278	0.847	0.838	0.825	0.871	**0.863**	0.853	-0.440
	XGBoost	0.079	0.066	0.052	0.321	0.293	0.277	0.849	0.830	0.797	0.864	0.845	0.815	-0.387
Córdoba (COR06)	BC		0.010			0.222			0.905			0.905		-0.786
	HS		0.016			0.246			0.884			0.885		-2.718
	ELM	0.009	0.003	0.000	0.232	0.211	0.202	0.921	0.914	0.897	0.921	0.915	0.898	0.156
	GRNN	0.027	0.020	0.015	0.239	0.228	0.219	0.908	0.900	0.890	0.908	0.901	0.891	-1.363
	MLP	-0.043	0.003	0.003	0.218	0.208	0.201	0.922	0.917	0.909	0.923	0.919	0.915	**0.431**
	RF	0.008	0.004	0.000	0.223	0.213	0.205	0.919	0.913	0.904	0.920	0.913	0.904	-0.027

	SVM	0.035	0.025	0.021	0.210	**0.206**	0.199	0.924	**0.918**	0.915	0.925	**0.920**	0.917	0.264
	XGBoost	0.009	**0.002**	0.002	0.251	0.222	0.211	0.914	0.905	0.879	0.914	0.905	0.881	-0.696
Loja (GRA03)	BC		0.189			0.295			0.786			0.886		-0.907
	HS		0.194			0.304			0.774			0.874		-1.387
	ELM	0.174	0.156	0.146	0.339	0.275	0.263	0.830	0.813	0.718	0.906	0.893	0.829	0.249
	GRNN	0.193	0.185	0.179	0.305	0.297	0.292	0.790	0.783	0.772	0.890	0.885	0.880	-0.927
	MLP	0.179	**0.154**	0.128	0.290	**0.267**	0.243	0.855	**0.825**	0.794	0.906	0.902	0.898	**0.677**
	RF	0.172	0.170	0.167	0.286	0.274	0.268	0.824	0.815	0.799	0.906	0.897	0.888	0.163
	SVM	0.201	0.183	0.170	0.305	0.287	0.279	0.808	0.797	0.771	0.908	**0.903**	0.900	-0.353
	XGBoost	0.181	0.161	0.145	0.294	0.280	0.271	0.819	0.808	0.788	0.903	0.890	0.880	0.018
El Campillo (HUE08)	BC		0.065			0.262			0.869			0.892		-1.139
	HS		0.086			0.291			0.839			0.859		-2.828
	ELM	0.093	0.083	0.077	0.236	0.227	0.220	0.907	0.902	0.894	0.922	0.917	0.908	0.203
	GRNN	0.114	0.107	0.102	0.293	0.258	0.242	0.887	0.873	0.836	0.911	0.896	0.864	-1.360
	MLP	0.129	**0.070**	0.030	0.243	**0.217**	0.202	0.921	**0.909**	0.887	0.927	**0.922**	0.916	**0.711**
	RF	0.090	0.085	0.080	0.246	0.232	0.226	0.902	0.897	0.885	0.919	0.912	0.900	-0.062
	SVM	0.115	0.103	0.099	0.244	0.232	0.227	0.901	0.897	0.886	0.923	0.920	0.914	-0.123
	XGBoost	0.091	0.078	0.061	0.256	0.239	0.228	0.901	0.890	0.875	0.914	0.904	0.891	-0.334
Mancha Real (JAE07)	BC		0.007			0.228			0.900			0.901		-0.976
	HS		0.019			0.249			0.882			0.882		-2.858
	ELM	0.013	0.008	0.005	0.217	0.209	0.202	0.922	0.916	0.910	0.922	0.917	0.910	0.538
	GRNN	0.034	0.027	0.020	0.242	0.231	0.223	0.905	0.898	0.888	0.907	0.900	0.891	-1.422
	MLP	0.043	**0.005**	0.000	0.218	**0.209**	0.199	0.924	**0.917**	0.909	0.925	**0.918**	0.910	**0.644**
	RF	0.013	0.009	0.006	0.232	0.220	0.211	0.915	0.907	0.897	0.915	0.907	0.897	-0.367
	SVM	0.040	0.030	0.027	0.222	**0.209**	0.201	0.923	**0.917**	0.906	0.924	**0.918**	0.909	0.288
	XGBoost	0.017	0.010	0.003	0.231	0.220	0.211	0.915	0.907	0.898	0.915	0.907	0.898	-0.387
Málaga (MAG01)	BC		0.030			0.283			0.826			0.828		-1.284
	HS		0.028			0.301			0.803			0.804		-2.790
	ELM	0.030	0.023	0.015	0.291	0.268	0.256	0.857	0.843	0.815	0.860	0.847	0.823	0.010

	GRNN	0.044	0.037	0.030	0.289	0.276	0.268	0.843	0.833	0.818	0.846	0.837	0.823	-0.801
	MLP	0.087	0.029	0.004	0.278	**0.262**	0.252	0.862	**0.850**	0.831	0.864	**0.857**	0.851	**0.480**
	RF	0.018	0.016	0.012	0.274	0.265	0.258	0.854	0.847	0.836	0.855	0.848	0.838	0.266
	SVM	0.065	0.054	0.048	0.273	0.267	0.261	0.851	0.844	0.837	0.864	0.855	0.848	-0.081
	XGBoost	0.028	**0.021**	0.007	0.277	0.271	0.264	0.848	0.839	0.833	0.851	0.842	0.836	-0.287
Écija (SEV09)	BC		0.085			0.225			0.897			0.915		-0.350
	HS		0.095			0.248			0.875			0.894		-2.186
	ELM	0.090	0.085	0.080	0.221	0.216	0.210	0.910	**0.905**	0.901	0.925	0.922	0.918	0.291
	GRNN	0.111	0.104	0.097	0.255	0.240	0.224	0.897	0.882	0.868	0.917	0.906	0.894	-1.555
	MLP	0.130	**0.084**	0.032	0.237	**0.216**	0.193	0.924	**0.905**	0.885	0.927	0.924	0.920	**0.367**
	RF	0.090	0.085	0.081	0.226	0.219	0.211	0.909	0.902	0.896	0.925	0.917	0.913	0.009
	SVM	0.105	0.101	0.098	0.225	0.220	0.214	0.907	0.901	0.896	0.929	**0.925**	0.922	0.004
	XGBoost	0.095	**0.084**	0.069	0.242	0.225	0.216	0.905	0.897	0.880	0.921	0.913	0.898	-0.392
Asheville (Ash08)	BC		0.010			0.300			0.741			0.738		-1.139
	HS		0.036			0.344			0.680			0.654		-2.566
	ELM	0.005	**0.002**	0.001	0.292	0.267	0.255	0.810	0.792	0.752	0.810	0.792	0.752	-0.204
	GRNN	0.022	0.014	0.003	0.266	0.249	0.236	0.838	0.821	0.799	0.837	0.819	0.794	0.175
	MLP	-0.067	-0.013	0.004	0.286	0.264	0.252	0.815	0.800	0.763	0.814	0.796	0.761	0.006
	RF	0.008	0.006	0.003	0.212	**0.199**	0.194	0.894	**0.887**	0.870	0.890	**0.884**	0.869	**1.442**
	SVM	0.026	0.016	0.013	0.287	0.266	0.257	0.808	0.794	0.761	0.807	0.793	0.760	-0.310
	XGBoost	0.006	0.003	0.000	0.279	0.230	0.187	0.897	0.844	0.774	0.897	0.844	0.773	0.706

4.2.Performance of the configurations

In order to assess the performance of the different configurations, Figure 4. 6 shows the mean values of RRMSE (a), NSE (b), MBE (c), and R^2 (d) for the different stations and models. Generally, on average, all configurations in Córdoba, Mancha Real, and Écija obtained the best performances, whereas, in Loja and Asheville, they had the worst results in most statistics (with the exceptions of Tabernas, Jimena de la Frontera, and Málaga in R^2). In terms of RRMSE, configuration 13 (light blue line) outperformed the rest in most locations (Tabernas, Córdoba, El Campillo, Mancha Real, and Málaga). However, some of them carried out better results using others, such as in Jimena de la Frontera using configurations 7 (pink circle) and 8 (black line), in Loja using configurations 8 (very close to 13), in Écija, using configuration 11 (blue line), and in Asheville with configurations 9 (red line), 7 (pink circle) and 14 (pink line). On the other hand, the less accurate results were obtained using configurations 6 and 11 in Asheville, configuration 10 (green line) in Tabernas, configuration 11 in Jimena de la Frontera, configuration 1 (black circle) in Córdoba, El Campillo, and Málaga, configuration 7 in Loja, and configuration 6 (light blue circle) in Mancha Real. Regarding NSE, the general performance was in line with RMSE, in which configurations 13, 8, and 11 had the highest values, while configurations 10 (Tabernas), 11 (Jimena de la Frontera and Asheville), 7 (Loja), 6 -light blue circle- (Mancha Real and Asheville), and 1 (Córdoba, El Campillo, Málaga, and Écija) were the worst. According to MBE, configurations 11 (Tabernas, El Campillo, Écija), 10 (Jimena de la Frontera, Mancha Real, and Málaga),9 (Asheville), 8 (Córdoba), and 5 -yellow circle- (Loja) outperformed the rest of the models, but on the other side, configurations 1 (Tabernas and Loja), 2 -red circle- (Mancha Real and Écija), 5 (Asheville), 6 (Jimena de la Frontera), 9 -red line- (Córdoba and Málaga) and 14 -pink line- (El Campillo) obtained the lowest improvements. Finally, in terms of R^2, configuration 13 outperformed in Tabernas, Córdoba, Loja, El Campillo, Mancha Real, and Málaga, and the results of configurations 9, 7 and 11 were the best ones for Asheville, Jimena de la Frontera and Écija, respectively.

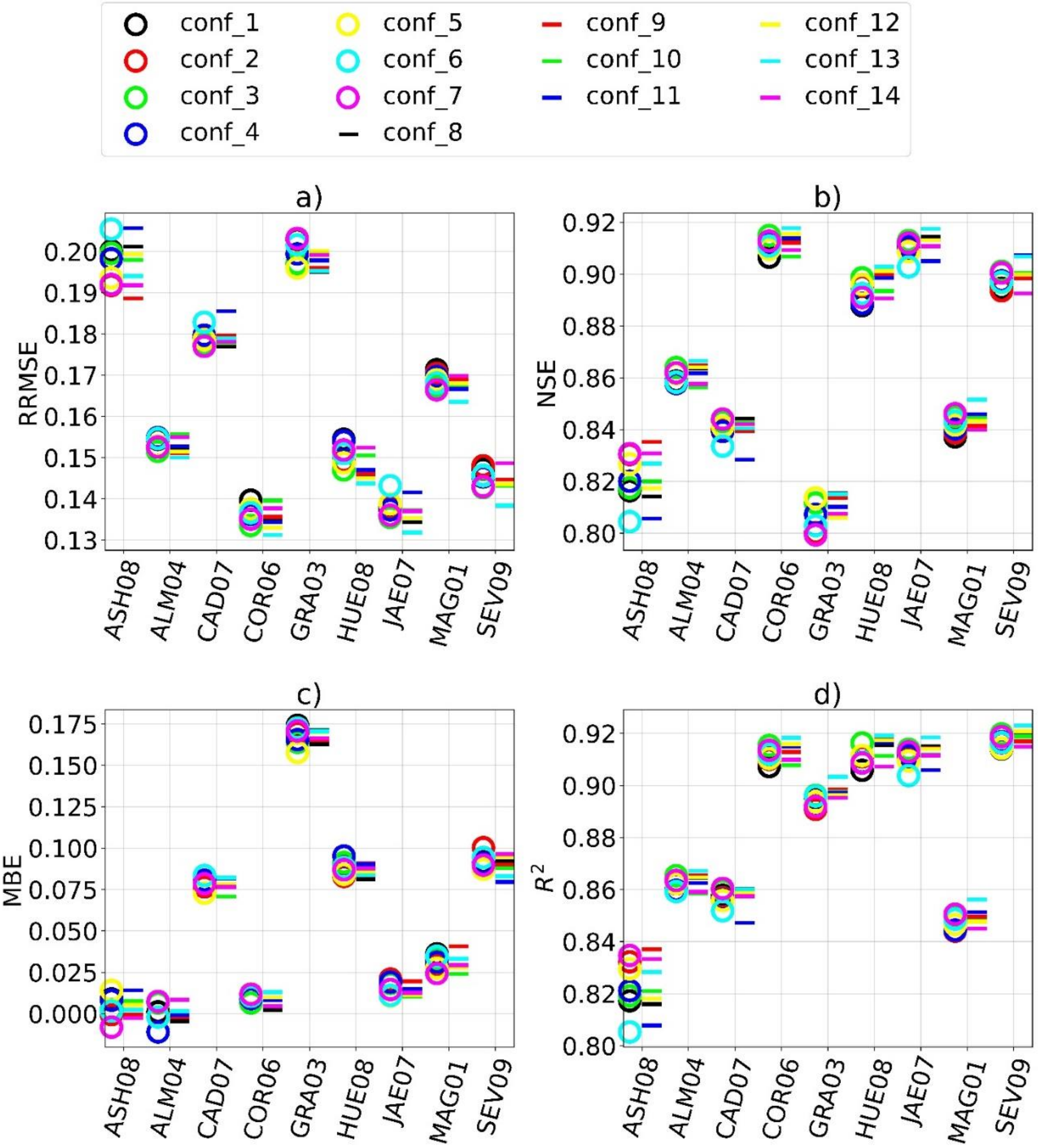


Figure 4. 6. Statistical performances for mean RRMSE (a), NSE (b), MBE (c) and R^2 (d) in all the stations assessed (ASH08 = Asheville, ALM04 = Tabernas, CAD07 = Jimena de la Frontera, COR06 = Córdoba, GRA03 = Loja, HUE08 = El Campillo, JAE07 = Mancha Real, MAG01 = Málaga, SEV09 = Écija) and using the different configuration

Table 4. 5 shows the best RMSE and NSE statistics, as well as the model used for each configuration and location, being the best values represented in bold. In Tabernas, configurations 13 (using ELM) and 5 (using MLP) obtained RMSE values below 0.24 Wm^{-2}. The NSE values ranged from NSE=0.862 (using configuration 1 and SVM) to NSE=0.874 (using configuration 13 and ELM). In Jimena de la Frontera, the three best configurations for both NSE and RMSE were configurations 8 (NSE=0.855 and RMSE=0.271 Wm^{-2} using MLP), 13 (NSE=0.855 and RMSE=0.272 Wm^{-2}), and 7 (NSE=0.853 and RMSE=0.273 Wm^{-2} using MLP), whereas configurations 6 (NSE=0.840 and RMSE=0.285 MJ m^{-2} using RF), 9 (NSE=0.844 and RMSE=0.282 Wm^{-2}) and 11

(NSE=0.844 and RMSE=0.281 Wm^{-2}) got the worst results. In Córdoba, the ranking for the 5 best configurations in terms of RMSE and NSE was: configuration 13 (NSE=0.924 and RMSE=0.199 Wm^{-2} using SVM), 7 (NSE=0.922 and RMSE=0.201 Wm^{-2} using MLP), 12 (NSE=0.921 and RMSE=0.202 Wm^{-2} using SVM), 9 (NSE=0.920 and RMSE=0.203 Wm^{-2} using SVM), 8 (NSE=0.920 and RMSE=0.204 Wm^{-2} using SVM). In Loja, MLP was the model used for the three best configurations: 8 (NSE=0.855 and RMSE=0.243 Wm^{-2}), 5 (NSE=0.842 and RMSE=0.253 Wm^{-2}), and 9 (NSE=0.840 and RMSE=0.255 Wm^{-2}), while the less promising configurations were 2 (NSE=0.814 and RMSE=0.275 Wm^{-2} using RF) and 12 (NSE=0.821 and RMSE=0.270 Wm^{-2} using ELM). In El Campillo, the NSE and RMSE ranged from NSE=0.921 and RMSE=0.202 Wm^{-2} using configuration 7, followed by configuration 13 (NSE=0.921 and RMSE=0.203 Wm^{-2}), to NSE=0.898 and RMSE=0.231 Wm^{-2} using configuration 4. In Mancha Real, the three configurations that outperformed the rest were 7 (NSE=0.924 and RMSE=0.199 Wm^{-2} using MLP), 13 (NSE=0.923 and RMSE=0.201 Wm^{-2} using MLP), and 12 (NSE=0.920 and RMSE=0.205 Wm^{-2} using SVM), whereas configurations 11 (NSE=0.910 and RMSE=0.217 Wm^{-2} using ELM) and 6 (NSE=0.910 and RMSE=0.217 Wm^{-2} using ELM) obtained the less promising results in this station. In Málaga, the NSE and RMSE values ranged from NSE=0.862 and RMSE=0.252 Wm^{-2} using configuration 13 and MLP, to NSE=0.846 and RMSE=0.266 MJ m^{-2} using configuration 1 and ELM. In Écija, the three best performances were carried out by configurations 13 (NSE=0.924 and RMSE=0.193 Wm^{-2}), 5 (NSE=0.914 and RMSE=0.205 Wm^{-2}), and 11 (NSE=0.913 and RMSE=0.207 Wm^{-2}) using MLP. In Asheville, the best statistics were obtained using configurations 9 (NSE=0.897 and RMSE=0.187Wm^{-2}), 2 (NSE=0.892 and RMSE=0.193 Wm^{-2}) and 5 (NSE=0.889 and RMSE=0.195 Wm^{-2}). Finally, attending to the global results for all the stations, the best configuration in terms of both statistics was 13, followed by 7 and 8.

Table 4. 5. Best NSE and RMSE (W/m^2) performances for the different configurations and stations assessed (ASH08 = Asheville, ALM04 = Tabernas. CAD07 = Jimena de la Frontera. COR06 = Córdoba. GRA03 = Loja. HUE08 = El Campillo. JAE07 = Mancha Real. MAG01 = Málaga. SEV09 = Écija) and the mean performance of all the stations. In parenthesis, it is shown the model used, while the best values are in bold.

	ALM04		**CAD07**		**COR06**		**GRA03**		**HUE08**		**JAE07**		**MAG01**		**SEV09**		**ASH08**		**All stations (mean)**	
Conf	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**	**NSE**	**RMSE**
1	0.862	0.248	0.847	0.279	0.916	0.209	0.835	0.259	0.908	0.219	0.917	0.208	0.846	0.266	0.907	0.213	0.882	0,201	0.880	0.234
	(SVM)	(SVM)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(ELM)	(ELM)	(MLP)	(MLP)	(RF)	(RF)		
2	0.867	0.244	0.850	0.275	0.918	0.207	0.814	0.275	0.914	0.212	0.916	0.209	0.850	0.262	0.904	0.217	0.892	0,193	0.881	0.233
	(MLP)	(MLP)	(RF)	(RF)	(MLP)	(MLP)	(RF)	(RF)	(MLP)	(MLP)	(ELM)	(ELM)	(MLP)	(MLP)	(ELM)	(ELM)	(XGB)	(XGB)		
3	0.868	0.242	0.850	0.276	0.919	0.205	0.838	0.256	0.907	0.220	0.919	0.206	0.850	0.262	0.906	0.214	0.887	0,197	0.883	0.231
	(SVM)	(SVM)	(RF)	(RF)	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(ELM)	(ELM)	(MLP)	(MLP)	(MLP)	(MLP)	(RF)	(RF)		
4	0.866	0.245	0.846	0.279	0.917	0.207	0.832	0.262	0.898	0.231	0.918	0.207	0.849	0.263	0.907	0.213	0.890	0,194	0.880	0.233
	(SVM)	(SVM)	(RF)	(RF)	(MLP)	(MLP)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(RF)	(RF)		
5	0.872	0.239	0.851	0.275	0.917	0.208	0.842	0.253	0.912	0.215	0.917	0.208	0.849	0.263	0.914	0.205	0.889	0,195	0.885	0.233
	(MLP)	(MLP)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(RF)	(RF)		
6	0.867	0.244	0.840	0.285	0.917	0.208	0.825	0.267	0.915	0.210	0.910	0.217	0.852	0.261	0.903	0.218	0.877	0,205	0.878	0.235
	(MLP)	(MLP)	(RF)	(RF)	(MLP)	(MLP)	(ELM)	(ELM)	(MLP)	(MLP)	(ELM)	(ELM)	(MLP)	(MLP)	(ELM)	(ELM)	(XGB)	(XGB)		
7	0.870	0.241	0.853	0.273	0.922	0.201	0.824	0.268	**0.921**	**0.202**	**0.924**	**0.199**	0.858	0.255	0.915	0.203	0.885	0,198	0.886	0.227
	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(RF)	(RF)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(RF)	(RF)		
8	0.869	0.242	**0.855**	**0.271**	0.920	0.204	**0.855**	**0.243**	0.919	0.206	0.919	0.205	0.852	0.260	0.906	0.215	0.884	0,199	0.887	0.227
	(SVM)	(SVM)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(MLP)	(MLP)	(RF)	(RF)	(RF)	(RF)		
9	0.870	0.241	0.844	0.282	0.920	0.203	0.840	0.255	0.906	0.221	0.918	0.207	0.847	0.264	0.904	0.217	**0.897**	**0,187**	0.883	0.231
	(SVM)	(SVM)	(RF)	(RF)	(SVM)	(SVM)	(MLP)	(MLP)	(ELM)	(ELM)	(SVM)	(SVM)	(ELM)	(ELM)	(ELM)	(ELM)	(XGB)	(XGB)		
10	0.865	0.245	0.848	0.278	0.917	0.207	0.822	0.269	0.905	0.222	0.918	0.207	0.855	0.258	0.909	0.211	0.888	0,196	0.881	0.233
	(SVM)	(SVM)	(RF)	(RF)	(SVM)	(SVM)	(ELM)	(ELM)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(RF)	(RF)		
11	0.865	0.245	0.844	0.281	0.917	0.208	0.837	0.258	0.909	0.218	0.910	0.217	0.851	0.261	0.913	0.207	0.871	0,210	0.880	0.234
	(ELM)	(ELM)	(RF)	(RF)	(ELM)	(ELM)	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(ELM)	(ELM)	(MLP)	(MLP)	(RF)	(RF)		
12	0.870	0.241	0.851	0.275	0.921	0.202	0.821	0.270	0.912	0.215	0.920	0.205	0.856	0.257	0.906	0.214	0.888	0,196	0.883	0.231
	(MLP)	(MLP)	(ELM)	(ELM)	(SVM)	(SVM)	(ELM)	(ELM)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(ELM)	(ELM)	(RF)	(RF)		
13	**0.874**	**0.237**	0.855	0.272	**0.924**	**0.199**	0.830	0.263	0.921	0.203	0.923	0.201	**0.862**	**0.252**	**0.924**	**0.193**	0.884	0,200	**0.889**	**0.224**
	(ELM)	(ELM)	(RF)	(RF)	(SVM)	(SVM)	(ELM)	(ELM)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(RF)	(RF)		
14	0.866	0.244	0.848	0.277	0.918	0.207	0.836	0.258	0.905	0.223	0.919	0.206	0.848	0.264	0.901	0.221	0.887	0,196	0.881	0.233
	(SVM)	(SVM)	(MLP)	(MLP)	(SVM)	(SVM)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(MLP)	(ELM)	(ELM)	(RF)	0,201		

In order to assess all the different configurations, Table 4. 6 shows a GPI evaluation for MBE, RMSE, RRMSE, $R^{2,}$ and NSE. In general, the worst configurations (lowest GPI values) were number 1 and 6, with a GPI below -0.3. On the other hand, the best ones were 13 and 8. In this sense, it is important to highlight the valuable information contained when using $Energy_T$ and $Hourmin_{Tx}$ in the same configuration because the worst results were performed by those which do not contain both. Moreover, the best configurations also incorporate $delta_T$ and temperature information about consecutive days (Tn_{prev}, $T_{x\ prev}$, and $T_{n\ next}$). In particular, the use of $delta_T$ is of high importance, despite being considered to be automatically learned by the model through T_x and T_n data. In addition, MLP, SVM, and ELM carried out the most accurate performances in most inland and medium aridity index locations and configurations for RMSE and NSE statistics (having ELM the lowest computational cost). However, in sub-humid and humid sites, the use of RF and XGBoost was better. Regarding the GPI evaluation for the different models, using MLP with configuration 13 was, in general, the most accurate model to estimate daily solar radiation.

Table 4. 6. GPI values for the different models (BC – Bristow and Campbell. HS – Hargreaves Samani. MLP – Multilayer Perceptron. ELM – Extreme Learning Machine. RF – Random Forest. SVM – Support Vector Machine. XGBoost – Extreme Gradient Boost. and GRNN – Generalized Regression Neural Network) and configurations (see table 39). The best values are in bold.

GPI for models		**GPI for configurations**			
Model	**GPI**	**Conf.**	**GPI**	**Conf.**	**GPI**
BC	-1.175	1	-0.371	9	0.007
HS	-2.593	2	-0.204	10	-0.165
MLP	**0.402**	3	0.048	11	-0.122
ELM	0.171	4	-0.220	12	0.030
RF	0.195	5	-0.128	**13**	**0.299**
SVM	-0.062	6	-0.355	14	-0.227
XGBoost	-0.221	7	-0.079		
GRNN	-1.069	8	0.075		

4.3.Seasonal performance

In order to evaluate the seasonal performance of the different models for all locations and configurations, Figure 4. 7 shows a boxplot of the statistics RRMSE (a), NSE (b), MBE (c), and R^2 (d) for winter, autumn, spring, and summer. In terms of RRMSE, the best mean performance was carried out in summer, while in winter and autumn, the results were worse. Regarding MBE, the highest variability performance was detected in spring and summer, ranging from around 0.3 MJ m^{-2} to -0.13 MJ m^{-2}, whereas winter performed consistently (all values are very close to the median). Moreover, all seasonal medians are positive in all models, which denoted a general underestimation trend. According to NSE and R^2, the highest and the lowest values, on average, were obtained in summer and autumn, respectively, being the seasonal performance quite similar in winter and spring.

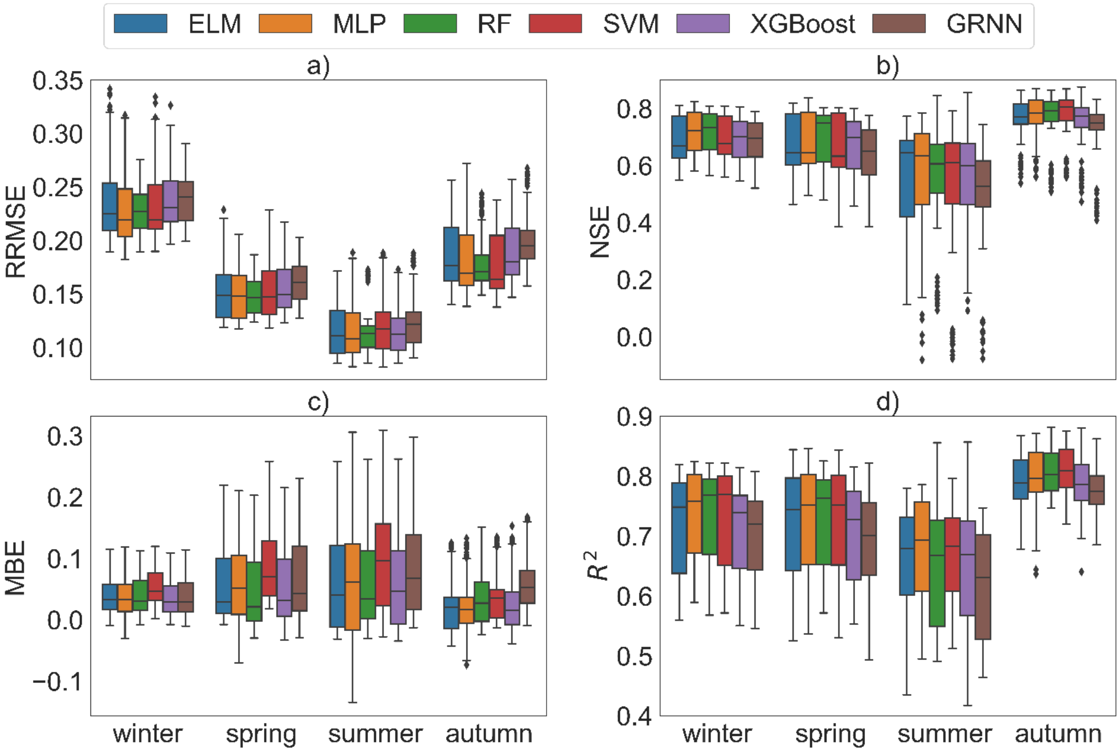


Figure 4. 7. Seasonal performance (RRMSE, NSE, MBE, and R^2) of the different models assessed (ELM – Extreme Learning Machine, RF – Random Forest, SVM – Support Vector Machine, XGB – Extreme Gradient Boost, MLP – Multilayer Perceptron, and GRNN – Generalized Regression Neural Network) in all the stations and configurations (Table 4. 3).

Table 4. 7 displays the mean seasonal NSE and RRMSE values of the different configurations for all the assessed stations and models, as well as the performance carried out by HS and BC. Configuration 13 (marked in bold) obtained the lowest RRMSE and the highest NSE in all seasons, with an improvement of 22.5% in summer against BC and 66.06% against HS. Generally, in terms of NSE, the most significant improvement was carried out in summer, although RRMSE had a general 8% improvement in all seasons, with slight variability in performance between different configurations.

Table 4. 7. Mean NSE and RRMSE seasonal values for the different configurations assessed (table 4), as well as HS (Hargreaves Samani) and BC (Bristow Campbell). The best values are in bold.

	NSE				**RRMSE**			
Conf.	**Winter**	**Spring**	**Summer**	**Autumn**	**Winter**	**Spring**	**Summer**	**Autumn**
HS	0.605	0.564	0.343	0.692	0.271	0.180	0.146	0.211
BC	0.643	0.621	0.462	0.723	0.255	0.167	0.131	0.198
1	0.690	0.658	0.541	0.759	0.239	0.157	0.122	0.186
2	0.693	0.674	0.541	0.757	0.237	0.154	0.121	0.187
3	0.702	0.682	0.558	0.760	0.234	0.152	0.119	0.186
4	0.694	0.673	0.541	0.758	0.238	0.154	0.122	0.187
5	0.696	0.675	0.548	0.762	0.235	0.153	0.118	0.184
6	0.689	0.660	0.537	0.754	0.238	0.156	0.121	0.188
7	0.695	0.675	0.535	0.766	0.231	0.151	0.117	0.180
8	0.705	0.682	0.551	0.763	0.232	0.151	0.119	0.184

9	0.704	0.684	0.552	0.761	0.232	0.150	0.118	0.185
10	0.697	0.675	0.543	0.752	0.235	0.153	0.120	0.188
11	0.700	0.669	0.548	0.758	0.234	0.154	0.120	0.186
12	0.702	0.680	0.547	0.764	0.233	0.152	0.120	0.183
13	**0.716**	**0.688**	**0.565**	**0.771**	**0.228**	**0.149**	**0.117**	**0.181**
14	0.696	0.672	0.536	0.760	0.235	0.154	0.121	0.185

To sum up, it could be stated that in autumn (followed by the results in winter), the best NSE values were obtained, whereas, in summer, the error was the highest for all configurations. However, in terms of RRMSE, the minimum error was given in summer, while the predictions in winter were the worst. This behavior is the consequence of the much higher solar radiation values in summer than in winter.

4.4. Models' application to new sites

To further assess the practical utility of these self-calibrated models for real applications, an evaluation during the same test period was carried out using non-used datasets from new sites with similar geo-climatic conditions (aridity index and coastal/inland locations). In these terms, Bélmez was tested with the best model and configuration from Córdoba (SVM using configuration 13), Almuñécar was assessed using the trained model from Málaga (MLP using configuration 7), Cabra was tested with El Campillo (using MLP and configuration 7), and finally, Huércal-Overa was evaluated with Tabernas (SVM using configuration 13). Eventually, these results were compared with those obtained using a non-calibrated HS daily prediction.

Table 4. 8. Statistical results of the models from Córdoba, El Campillo, Málaga and Tabernas for new non-trained sites with similar climatic conditions (Bélmez, Cabra, Almuñécar and Huércal-Overa). as well as the non-calibrated Hargreaves-Samani results.

		Non-calibrated Hargreaves-Samani				**Machine learning model**			
Model from	**Test Station**	**RMSE** [Wm^2]	**MBE** [Wm^2]	**R^2**	**NSE**	**RMSE** [Wm^2]	**MBE** [Wm^2]	**R^2**	**NSE**
Córdoba	Bélmez	0.273	-0.057	0.880	0.871	0.217	0.037	0.915	0.912
El Campillo	Cabra	0.344	-0.163	0.849	0.786	0.233	-0.036	0.903	0.900
Málaga	Almuñécar	0.305	0.134	0.853	0.818	0.252	0.078	0.8756	0.862
Tabernas	Huércal-Overa	0.311	-0.097	0.836	0.802	0.249	0.013	0.873	0.872

Table 4. 8 shows the statistical performance of both HS and ML for these new sites. In general, all ML models highly outperformed non-calibrated HS. In Bélmez, with an aridity index of 0.3681 (semiarid location), the RMSE values were reduced from 0.273 Wm^{-2} to 0.217 Wm^{-2} and improved NSE from 0.871 to 0.912. In Cabra, with an aridity index of 0.5641, the R^2 was outperformed from 0.849 using HS to 0.903 using ML and reducing RMSE and MBE values from RMSE=0.344 Wm^{-2} MBE=-0.163 Wm^{-2} to RMSE=0.233 Wm^{-2} and MBE=-0.036Wm^{-2} for HS and ML, respectively. In Huércal-Overa, with an aridity index of 0.2284, the NSE was outperformed from 0.802 to 0.872 for HS and ML, respectively. The RMSE was reduced from RMSE=0.311 Wm^{-2} to RMSE 0.249 Wm^{-2}. Finally, in the coastal location of Almuñécar, the RMSE and NSE values carried out by HS were RMSE=0.305 Wm^{-2} and NSE= 0.818 Wm^{-2}, against the performance of ML of RMSE=0.252 Wm^{-2} and NSE=0.862. Another substantial

reduction was seen in MBE, from MBE=0.134 Wm^{-2} to MBE=0.078 Wm^{-2}. In all cases, the percentage of improvement in the evaluated statistics ranged from 17.3% to 32.2% for RMSE, from 41.7% to 164.9% for MBE, from 2.5% to 6.3% for R^2, and from 4.7% to 14.5% for NSE. Thus, the results obtained corroborate the extrapolation ability of the proposed models using these novel configurations when applying them to new sites.

4.5.Overall discussion

The results obtained by these ML models outperformed those given by Estévez et al. [22] for inland locations (RMSE=0.283 Wm^{-2}) and also for coastal sites (RMSE=0.312 Wm^{-2}) in this same region. In terms of R^2 and RRMSE, these models got better results than those reported by Feng et al. [37] in four arid stations assessed in China (Turpan, Yinchuan, Dunhuang, and Xilingol) using ANN (Artificial Neural Network), MEA-ANN (Mind Evolutionary Algorithm Artificial Neural Network), RF and WNN (Wavelet Neural Network) with an R^2 range from 0.707 in Yinchuan using RF, to 0.885 in Turpan using MEA-ANN and an RRMSE range from 0.191 in Dunhuang using WNN, to 0.251 in Yinchuan using WNN. Furthermore, it also improved the results of Fan et al. [93] in three humid subtropical climate stations (Wuhan, Kunming, and Guangzhou) using SVM and XGBoost in terms of R^2 and RMSE, where the best R^2 were carried out by SVM in Wuhan (R^2=0.776) and the best RMSE in Guangzhou also using SVM (RMSE= 0,243 W m^{-2}). In terms of R^2, this study's results also improved the performances of Sharifi et al. [38] for the six studied stations in Iran and the United States using ANN, GEP, and WR, ranging from R^2=0.733 in Lincoln with GEP to R^2=0.875 in Urmia with ANN. Comparing the results with the work of Jiménez et al. [94], Neural Network's performance also using the relative humidity the pressure is slightly better in terms of RRMSE, ranging from RRMSE=0.0811 in El Colmenar to RRMSE=0.1006 in Santa Ana. Finally, comparing the results with the work of Bellido et al. [89], predicting solar radiation using CNN and LSTM with hourly temperature in Almuñecar, the results are highly outperformed with the approach followed in this work.

In general, no model predominantly outperformed the rest, although MLP, SVM, and ELM were usually among the best approaches for semiarid and arid sites, whereas RF and XGBoost obtained better performances in sub-humid and humid locations. In order to obtain the best model (using all the configurations) and the best configuration (for all the models studied), the corresponding GPI values were calculated (Table 4. 6). As it can be seen marked in bold, the use of the MLP model and configuration 13 (T_X, Tn, R_a, DTR, $Energy_T$, $Hourmin_{Tx}$, $T_{x\ prev,}$ and $T_{n\ next}$) was preferred to generally estimate daily solar radiation values due to its better performance. Additionally, based on the results of this study, and despite $delta_T$ is a piece of information that could be obtained from T_x and T_n during the iteratively training process, its use as an input variable helps ML models to estimate daily solar radiation more accurately, as well as the introduction of new variables based on intra-daily temperatures like $Energy_T$ and $Hourmin_{T_x}$ [82]. They help the models generalize better for their use in different locations from the trained. Likewise, this work also highlights the importance of the $T_{X\ prev}$ and the Tn_{next} (as it was stated by Bristow and Campbell [60]) due to the effect of advection cold (or warm) air masses in $delta_T$. Thus, the contribution of the new temperature-based variables used is crucial for improving daily solar radiation predictions using ML models, being a meaningful way to deepen future works worldwide, mainly due to its easy implementation and low-cost measurement.

5. Conclusions and future work

Several locations representing different geo-climatic conditions were studied in order to assess several machine learning models for daily solar radiation predictions using different intra-daily temperature-based variables. Firstly, the period from 2000 to 2014 was used for training purposes and tuned the different hyperparameters. Afterward, its performance was evaluated using the period from 2014 to 2018. The main novelties of this study are the use of new air temperature variables to estimate solar radiation like $Energy_T$, $Hourmin_{Tx}$, $Hourmin_{Tx\ prev}$, and the application of Bayesian Optimization instead of the commonly used trial and error technique. The results arose a good behavior, especially with the general use of MLP and configuration 13 (T_X, Tn, R_a, DTR, $Energy_T$, $Hourmin_{Tx}$, $T_{X\ prev,}$ and Tn_{next}), outperforming the results of the empirical and widely used methods HS and BC in all the assessed stations. However, a more specific use of MLP and SVM for semiarid and arid sites and Rf and XGBoost for dry sub-humid and humid sites is preferred. Furthermore, on average, all configurations and all the models evaluated except GRNN improved the results from empirical models and their excellent performances in different geo-climatic conditions. For example, in terms of NSE: the most significant improvement was from 0.741 using BC to 0.897 using RF in the most humid location, and even in a semiarid inland site with adequate empirical performance (NSE=0.905 using BC), the use of SVM was able to achieve a value of NSE=0.924; in terms of RMSE: comparing the results to BC and HS, the improvement ranged from 7.56% in the aridest station using ELM to 45.65% in a humid location using MLP and in terms of mean GPI values: the most accurate models were MLP for locations with medium aridity values (between 0.2808 and 0.5497), SVM for the aridest site and RF for the most humid site. Moreover, using MLP and configuration 13, there was a mean error reduction of around 8% per season in terms of RRMSE and up to 60% (for HS in summer) in terms of seasonal NSE.

One of the common problems when using local-calibrated methods is the prerequisite of using enough datasets for training and testing processes at each site. Several models have been evaluated in new locations with similar geo-climatic characteristics to the trained sites in this work. The results highly outperformed HS in all conditions, with an improvement of up to 14.5% in terms of NSE and 47.6% in terms of RMSE in dry sub-humid climate sites. Thus, using these models could improve solar radiation predictions at new sites only using air temperature measurements, although a previously trained model (with similar geo-climatic features) is required for this purpose.

However, further studies can evaluate the efficiency of these models/configurations in other new areas, as well as the application on a regional scenario, due to the excellent performance of the models when applied to new datasets from different (but with similar geo-climatic conditions) sites. Finally, due to the increasing use of low-cost temperature sensors and the expansion of IoT devices, these new ML approaches may result in more accurate solar radiation predictions at a great variety of locations where there is no possibility of installing solar radiation sensors (regions with lack of funding, developing countries, among others), as well as optimizing the determination of potential locations for solar power plants construction. Furthermore, the combination of the air temperature forecasts from climate change (or conventional numerical weather prediction) models, the developed approaches of this work could be of great interest for the potential efficiency of solar energy systems in the medium-large term.

6. References

[1] Masson-Delmotte V, Zhai P, Pörtner H-O, Roberts D, Skea J, Calvo E, et al. IPCC, 2019: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food 2019.

[2] Wuebbles DJ, Fahey DW, Hibbard KA. Climate science special report: fourth national climate assessment, vol. I 2017.

[3] Chang W, Stein ML, Wang J, Kotamarthi VR, Moyer EJ. Changes in spatiotemporal precipitation patterns in changing climate conditions. J Clim 2016;29:8355–76. https://doi.org/10.1175/JCLI-D-15-0844.1.

[4] Sohoulande Djebou DC, Singh VP. Impact of climate change on precipitation patterns: a comparative approach. International Journal of Climatology 2016;36:3588–606. https://doi.org/10.1002/JOC.4578.

[5] Portner H-O, Roberts DC, Masson-Delmotte V, Zhai P, Tignor M, Poloczanska E, et al. PCC 2019: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate 2019. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=PCC+2019%3A+IPCC+Special+Report+on+the+Ocean+and+Cryosphere+in+a+Changing+Climate&btnG= (accessed March 20, 2023).

[6] Knutson TR, McBride JL, Chan J, Emanuel K, Holland G, Landsea C, et al. Tropical cyclones and climate change. Nat Geosci 2010;3:157–63. https://doi.org/10.1038/NGEO779.

[7] Estévez J, Gavilán P, Giráldez J V. Guidelines on validation procedures for meteorological data from automatic weather stations. J Hydrol (Amst) 2011;402:144–54. https://doi.org/10.1016/j.jhydrol.2011.02.031.

[8] Hosseini-Fashami F, Motevali A, Nabavi-Pelesaraei A, Hashemi SJ, Chau K wing. Energy-Life cycle assessment on applying solar technologies for greenhouse strawberry production. Renewable and Sustainable Energy Reviews 2019;116. https://doi.org/10.1016/j.rser.2019.109411.

[9] Ghasemi-Mobtaker H, Mostashari-Rad F, Saber Z, Chau K wing, Nabavi-Pelesaraei A. Application of photovoltaic system to modify energy use, environmental damages and cumulative exergy demand of two irrigation systems-A

case study: Barley production of Iran. Renew Energy 2020;160:1316–34. https://doi.org/10.1016/j.renene.2020.07.047.

[10] Singh GK. Solar power generation by PV (photovoltaic) technology: A review. Energy 2013;53:1–13. https://doi.org/10.1016/j.energy.2013.02.057.

[11] Jager-Waldau A, Adinolfi G, Batlle A, Braun M, Bucher C, Detollenaere A, et al. Electricity produced from photovoltaic systems in apartment buildings and self-consumption : Comparison of the situation in various IEA PVPS countries. Conference Record of the IEEE Photovoltaic Specialists Conference 2019:1701–10. https://doi.org/10.1109/PVSC40753.2019.8980484.

[12] Jäger-Waldau. Pv status report 2019 - Publications Office of the EU. European Commission, Joint Research Centre 2016.

[13] Jäger-Waldau A. Snapshot of Photovoltaics—February 2020. Energies 2020, Vol 13, Page 930 2020;13:930. https://doi.org/10.3390/EN13040930.

[14] Sánchez-Lozano JM, Teruel-Solano J, Soto-Elvira PL, Socorro García-Cascales M. Geographical Information Systems (GIS) and Multi-Criteria Decision Making (MCDM) methods for the evaluation of solar farms locations: Case study in south-eastern Spain. Renewable and Sustainable Energy Reviews 2013;24:544–56. https://doi.org/10.1016/j.rser.2013.03.019.

[15] André C, Jonathan M, César L, Carlos H. Use of Renewable Energies in Sustainable Facilities for Rural Housing Module. MATEC Web of Conferences 2019;303:02002. https://doi.org/10.1051/MATECCONF/201930302002.

[16] Gašparović I, Gašparović M. Determining optimal solar power plant locations based on remote sensing and GIS methods: A case study from Croatia. Remote Sens (Basel) 2019;11. https://doi.org/10.3390/RS11121481.

[17] Thevenard D. REVIEW AND RECOMMENDATIONS FOR IMPROVING THE MODELLING OF BUILDING INTEGRATED PHOTOVOLTAIC SYSTEMS n.d.

[18] Smirnov V, Lambertz A, Böttler W, Carius R, Finger F, Hüpkes J, et al. Overview of Temperature Coefficients of Different Thin Film Photovoltaic Technologies. 25th European Photovoltaic Solar Energy Conference and Exhibition / 5th World Conference on Photovoltaic Energy Conversion, 6-10 September 2010, Valencia, Spain 2010:4248–52. https://doi.org/10.4229/25THEUPVSEC2010-4AV.3.83.

[19] Janjai S, Pankaew P, Laksanaboonsong J. A model for calculating hourly global solar radiation from satellite data in the tropics. Appl Energy 2009;86:1450–7. https://doi.org/10.1016/j.apenergy.2009.02.005.

[20] Ghimire S, Deo RC, Raj N, Mi J. Deep learning neural networks trained with MODIS satellite-derived predictors for long-term global solar radiation prediction. Energies (Basel) 2019;12. https://doi.org/10.3390/EN12122407.

[21] Amente G., Abate Y. Comparison of Different Empirical Models In The Estimatio... - Google Scholar. Ethiopia: Haramaya University 2014.

[22] Estévez J, Padilla FLM, Gavilán P. Evaluation and Regional Calibration of Solar Radiation Prediction Models in Southern Spain. Journal of Irrigation and Drainage Engineering 2012;138:868–79. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000493.

[23] Yagli GM, Yang D, Srinivasan D. Automatic hourly solar forecasting using machine learning models. Renewable and Sustainable Energy Reviews 2019;105:487–98. https://doi.org/10.1016/j.rser.2019.02.006.

[24] Trnka M, Žalud Z, Eitzinger J, Dubrovský M. Global solar radiation in Central European lowlands estimated by various empirical formulae. Agric For Meteorol 2005;131:54–76. https://doi.org/10.1016/j.agrformet.2005.05.002.

[25] Chen JL, He L, Yang H, Ma M, Chen Q, Wu SJ, et al. Empirical models for estimating monthly global solar radiation: A most comprehensive review and comparative case study in China. Renewable and Sustainable Energy Reviews 2019;108:91–111. https://doi.org/10.1016/j.rser.2019.03.033.

[26] Naserpour S, Zolfaghari H, Zeaiean Firouzabadi P. Calibration and evaluation of sunshine-based empirical models for estimating daily solar radiation in Iran. Sustainable Energy Technologies and Assessments 2020;42. https://doi.org/10.1016/j.seta.2020.100855.

[27] Hassan GE, Youssef ME, Mohamed ZE, Ali MA, Hanafy AA. New Temperature-based Models for Predicting Global Solar Radiation. Appl Energy 2016;179:437–50. https://doi.org/10.1016/j.apenergy.2016.07.006.

[28] Wazira Azhari A, Sopian K, Zaharim A, Ghoul M AL. A New Approach For Predicting Solar Radiation In Tropical Environment Using Satellite Images-Case Study Of Malaysia n.d.

[29] Fu C-L, Cheng H-Y. Predicting solar irradiance with all-sky image features via regression. Solar Energy 2013;97:537–50. https://doi.org/10.1016/j.solener.2013.09.016.

[30] Liu Y, Zhou Y, Chen Y, Wang D, Wang Y, Zhu Y. Comparison of support vector machine and copula-based nonlinear quantile regression for estimating the daily diffuse solar radiation: A case study in China. Renew Energy 2020;146:1101–12. https://doi.org/10.1016/j.renene.2019.07.053.

[31] Ghimire S, Deo RC, Raj N, Mi J. Wavelet-based 3-phase hybrid SVR model trained with satellite-derived predictors, particle swarm optimization and

maximum overlap discrete wavelet transform for solar radiation prediction. Renewable and Sustainable Energy Reviews 2019;113. https://doi.org/10.1016/j.rser.2019.109247.

[32] Srivastava S, Lessmann S. A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data. Solar Energy 2018;162:232–47. https://doi.org/10.1016/j.solener.2018.01.005.

[33] Amiri B, Gómez-Orellana AM, Gutiérrez PA, Dizène R, Hervás-Martínez C, Dahmani K. A novel approach for global solar irradiation forecasting on tilted plane using Hybrid Evolutionary Neural Networks. J Clean Prod 2021;287. https://doi.org/10.1016/j.jclepro.2020.125577.

[34] Guijo-Rubio D, Durán-Rosal AM, Gutiérrez PA, Gómez-Orellana AM, Casanova-Mateo C, Sanz-Justo J, et al. Evolutionary artificial neural networks for accurate solar radiation prediction. Energy 2020;210. https://doi.org/10.1016/j.energy.2020.118374.

[35] Tao H, Ewees AA, Al-Sulttani AO, Beyaztas U, Hameed MM, Salih SQ, et al. Global solar radiation prediction over North Dakota using air temperature: Development of novel hybrid intelligence model. Energy Reports 2021;7:136–57. https://doi.org/10.1016/j.egyr.2020.11.033.

[36] Pang Z, Niu F, O'Neill Z. Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons. Renew Energy 2020;156:279–89. https://doi.org/10.1016/j.renene.2020.04.042.

[37] Feng Y, Gong D, Zhang Q, Jiang S, Zhao L, Cui N. Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. Energy Convers Manag 2019;198. https://doi.org/10.1016/j.enconman.2019.111780.

[38] Sharifi SS, Rezaverdinejad V, Nourani V. Estimation of daily global solar radiation using wavelet regression, ANN, GEP and empirical models: A comparative study of selected temperature-based approaches. J Atmos Sol Terr Phys 2016;149:131–45. https://doi.org/10.1016/J.JASTP.2016.10.008.

[39] Jimenez VA, Barrionuevo A, Will A, Rodríguez S. Neural Network for Estimating Daily Global Solar Radiation Using Temperature, Humidity and Pressure as Unique Climatic Input Variables. Smart Grid and Renewable Energy 2016;07:94–103. https://doi.org/10.4236/SGRE.2016.73006.

[40] Davcev D, Mitreski K, Trajkovic S, Nikolovski V, Koteli N. IoT agriculture system based on LoRaWAN. IEEE International Workshop on Factory Communication Systems - Proceedings, WFCS 2018;2018-June:1–4. https://doi.org/10.1109/WFCS.2018.8402368.

[41] Luna AM, Lineros ML, Gualda JE, Cervera JVG, Luna JMM. Assessing the Best Gap-Filling Technique for River Stage Data Suitable for Low Capacity

Processors and Real-Time Application Using IoT. Sensors 2020, Vol 20, Page 6354 2020;20:6354. https://doi.org/10.3390/S20216354.

[42] Mesas-Carrascosa FJ, Verdú Santano D, Meroño JE, Sánchez de la Orden M, García-Ferrer A. Open source hardware to monitor environmental parameters in precision agriculture. Biosyst Eng 2015;137:73–83. https://doi.org/10.1016/J.BIOSYSTEMSENG.2015.07.005.

[43] Nsabagwa M, Byamukama M, Kondela E, Otim JS. Towards a robust and affordable Automatic Weather Station. Dev Eng 2019;4. https://doi.org/10.1016/J.DEVENG.2018.100040.

[44] Strigaro D, Cannata M, Antonovic M. Boosting a weather monitoring system in low income economies using open and non-conventional systems: Data quality analysis. Sensors (Switzerland) 2019;19. https://doi.org/10.3390/S19051185.

[45] Zanetti SS, Dohler RE, Cecílio RA, Pezzopane JEM, Xavier AC. Proposal for the use of daily thermal amplitude for the calibration of the Hargreaves-Samani equation. J Hydrol (Amst) 2019;571:193–201. https://doi.org/10.1016/J.JHYDROL.2019.01.049.

[46] Ferreira LB, Da Cunha FF, Duarte AB, Sediyama GC, Cecon PR. Calibration methods for the Hargreaves-Samani equation. Ciencia e Agrotecnologia 2018;42:104–14. https://doi.org/10.1590/1413-70542018421017517.

[47] Camayo-Lapa DrBF, Pomachagua-Paucar DrJE, Massipe-Hernández DrJR, Quispe-Flores DrMO, Torres-Ten MScA. Validation and application of Bristow Campbell model for estimating the global solar radiation in the Junin region. Tecnologia Quimica 2017;37:574–90.

[48] Deo RC, Şahin M, Adamowski JF, Mi J. Universally deployable extreme learning machines integrated with remotely sensed MODIS satellite predictors over Australia to forecast global solar radiation: A new approach. Renewable and Sustainable Energy Reviews 2019;104:235–61. https://doi.org/10.1016/J.RSER.2019.01.009.

[49] Behera MK, Majumder I, Nayak N. Solar photovoltaic power forecasting using optimized modified extreme learning machine technique. Engineering Science and Technology, an International Journal 2018;21:428–38. https://doi.org/10.1016/J.JESTCH.2018.04.013.

[50] Lotfinejad MM, Hafezi R, Khanali M, Hosseini SS, Mehrpooya M, Shamshirband S. A comparative assessment of predicting daily solar radiation using Bat Neural Network (BNN), Generalized Regression Neural Network (GRNN), and Neuro-Fuzzy (NF) system: A case study. Energies (Basel) 2018;11. https://doi.org/10.3390/EN11051188.

[51] Guermoui M, Gairaa K, Boland J, Arrif T. A Novel Hybrid Model for Solar Radiation Forecasting Using Support Vector Machine and Bee Colony

Optimization Algorithm: Review and Case Study. Journal of Solar Energy Engineering, Transactions of the ASME 2021;143. https://doi.org/10.1115/1.4047852.

[52] Fan J, Wu L, Zhang F, Cai H, Wang X, Lu X, et al. Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature. Renewable and Sustainable Energy Reviews 2018;94:732–47. https://doi.org/10.1016/J.RSER.2018.06.029.

[53] Karasu S, Altan A. Recognition Model for Solar Radiation Time Series based on Random Forest with Feature Selection Approach. ELECO 2019 - 11th International Conference on Electrical and Electronics Engineering 2019:8–11. https://doi.org/10.23919/ELECO47770.2019.8990664.

[54] Ibrahim IA, Khatib T. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers Manag 2017;138:413–25. https://doi.org/10.1016/J.ENCONMAN.2017.02.006.

[55] Ma Z, Chang H, Sun Z, Liu F, Li W, Zhao D, et al. Very Short-Term Renewable Energy Power Prediction Using XGBoost Optimized by TPE Algorithm. 2020 4th International Conference on HVDC, HVDC 2020 2020:1236–41. https://doi.org/10.1109/HVDC50696.2020.9292870.

[56] Iizumi T, Nishimori M, Yokozawa M. Evaporation from a water surface in relation to solar radiation. Trans Roy Soc S Aust 1940;46:114–8. https://doi.org/10.2480/AGRMET.64.9.

[57] Angstrom A. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Quarterly Journal of the Royal Meteorological Society 1924;50:121–6. https://doi.org/10.1002/QJ.49705021008.

[58] Hargreaves GH, Samani ZA. Estimating Potential Evapotranspiration. Journal of the Irrigation and Drainage Division 1982;108:225–30. https://doi.org/10.1061/JRCEA4.0001390.

[59] Hargreaves GH, Samani ZA. Reference Crop Evapotranspiration from Temperature. Appl Eng Agric 1985;1:96–9. https://doi.org/10.13031/2013.26773.

[60] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric For Meteorol 1984;31:159–66. https://doi.org/10.1016/0168-1923(84)90017-0.

[61] Rivero M, Orozco S, Sellschopp FS, Loera-Palomo R. A new methodology to extend the validity of the Hargreaves-Samani model to estimate global solar radiation in different climates: Case study Mexico. Renew Energy 2017;114:1340–52. https://doi.org/10.1016/J.RENENE.2017.08.003.

[62] Pandey V, Pandey PK, Mahanta AP. Calibration and performance verification of hargreaves samani equation in a humid region. Irrigation and Drainage 2014;63:659–67. https://doi.org/10.1002/IRD.1874.

[63] Pan T, Wu S, Dai E, Liu Y. Estimating the daily global solar radiation spatial distribution from diurnal temperature ranges over the tibetan plateau in China. Appl Energy 2013;107:384–93. https://doi.org/10.1016/J.APENERGY.2013.02.053.

[64] Lee J, Kim CG, Lee JE, Kim NW, Kim H. Application of artificial neural networks to rainfall forecasting in the Geum River Basin, Korea. Water (Switzerland) 2018;10. https://doi.org/10.3390/W10101448.

[65] Abbot J, Marohasy J. Forecasting of Medium-term Rainfall Using Artificial Neural Networks: Case Studies from Eastern Australia. Engineering and Mathematical Topics in Rainfall 2018. https://doi.org/10.5772/INTECHOPEN.72619.

[66] Estévez J, Bellido-Jiménez JA, Liu X, García-Marín AP. Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment. Water (Switzerland) 2020;12. https://doi.org/10.3390/W12071909.

[67] Salcedo-Sanz S, Deo RC, Carro-Calvo L, Saavedra-Moreno B. Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. Theor Appl Climatol 2016;125:13–25. https://doi.org/10.1007/S00704-015-1480-4.

[68] Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. J Hydrol (Amst) 2019;572:556–70. https://doi.org/10.1016/J.JHYDROL.2019.03.028.

[69] Traore S, Luo Y, Fipps G. Deployment of artificial neural network for short-term forecasting of evapotranspiration using public weather forecast restricted messages. Agric Water Manag 2016;163:363–79. https://doi.org/10.1016/J.AGWAT.2015.10.009.

[70] Huang G Bin, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. Neurocomputing 2006;70:489–501. https://doi.org/10.1016/J.NEUCOM.2005.12.126.

[71] Feng Y, Peng Y, Cui N, Gong D, Zhang K. Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. Comput Electron Agric 2017;136:71–8. https://doi.org/10.1016/J.COMPAG.2017.01.027.

[72] Şahin M, Kaya Y, Uyar M, Yildirim S. Application of extreme learning machine for estimating solar radiation from satellite data. Int J Energy Res 2014;38:205–12. https://doi.org/10.1002/ER.3030.

[73] Yin H, Dong Z, Chen Y, Ge J, Lai LL, Vaccaro A, et al. An effective secondary decomposition approach for wind power forecasting using extreme learning machine trained by crisscross optimization. Energy Convers Manag 2017;150:108–21. https://doi.org/10.1016/J.ENCONMAN.2017.08.014.

[74] Ahmad W, Ayub N, Ali T, Irfan M, Awais M, Shiraz M, et al. Towards short term electricity load forecasting using improved support vector machine and extreme learning machine. Energies (Basel) 2020;13. https://doi.org/10.3390/EN13112907.

[75] Ghritlahre HK, Prasad RK. Exergetic performance prediction of solar air heater using MLP, GRNN and RBF models of artificial neural network technique. J Environ Manage 2018;223:566–75. https://doi.org/10.1016/J.JENVMAN.2018.06.033.

[76] AL-Musaylh MS, Deo RC, Li Y, Adamowski JF. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. Appl Energy 2018;217:422–39. https://doi.org/10.1016/J.APENERGY.2018.02.140.

[77] Mohammadi K, Shamshirband S, Anisi MH, Amjad Alam K, Petković D. Support vector regression based prediction of global solar radiation on a horizontal surface. Energy Convers Manag 2015;91:433–41. https://doi.org/10.1016/J.ENCONMAN.2014.12.015.

[78] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[79] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[80] Chen T, He T. xgboost: eXtreme Gradient Boosting n.d.

[81] Middleton N, Thomas DSG. World atlas of desertification. World Atlas of Desertification 1992. https://doi.org/10.2307/3060449.

[82] Bellido-Jiménez JA, Estévez J, García-Marín AP. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agricultural Water Management 2020:106558. https://doi.org/10.1016/j.agwat.2020.106558.

[83] Estévez J, Gavilán P, García-Marín AP. Spatial regression test for ensuring temperature data quality in southern Spain. Theor Appl Climatol 2018;131:309–18. https://doi.org/10.1007/S00704-016-1982-8.

[84] Estévez J, Gavilán P, García-Marín AP. Spatial regression test for ensuring temperature data quality in southern Spain. Theoretical and Applied Climatology 2018;131:309–18. https://doi.org/10.1007/s00704-016-1982-8.

[85] Wang K, Qi X, Liu H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. Appl Energy 2019;251. https://doi.org/10.1016/J.APENERGY.2019.113315.

[86] Torres JF, Troncoso A, Koprinska I, Wang Z, Martínez-Álvarez F. Deep Learning for Big Data Time Series Forecasting Applied to Solar Power. Advances in Intelligent Systems and Computing 2019;771:123–33. https://doi.org/10.1007/978-3-319-94120-2_12/COVER.

[87] Močkus J. On Bayesian Methods for Seeking the Extremum. Optimization Techniques IFIP Technical Conference 1975:400–4. https://doi.org/10.1007/978-3-662-38527-2_55.

[88] Bellido-Jiménez JA, Estévez J, García-Marín AP. Assessing Neural Network Approaches for Solar Radiation Estimates Using Limited Climatic Data in the Mediterranean Sea. Environmental Sciences Proceedings 2021, Vol 4, Page 19 2020;4:19. https://doi.org/10.3390/ECAS2020-08116.

[89] Bellido-Jiménez JA, Estévez J, García-Marín AP. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agric Water Manag 2021;245. https://doi.org/10.1016/j.agwat.2020.106558.

[90] Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. Adv Neural Inf Process Syst 2012;25.

[91] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.

[92] Despotovic M, Nedic V, Despotovic D, Cvetanovic S. Review and statistical analysis of different global solar radiation sunshine models. Renewable and Sustainable Energy Reviews 2015;52:1869–80. https://doi.org/10.1016/J.RSER.2015.08.035.

[93] Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, et al. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy Convers Manag 2018;164:102–11. https://doi.org/10.1016/J.ENCONMAN.2018.02.087.

[94] Jimenez VA, Barrionuevo A, Will A, Rodríguez S. Neural network for estimating daily global solar radiation using temperature, humidity and pressure as unique climatic input variables. Smart Grid Renewable Energy 2016;7:94.

Chapter 5

Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain

Juan Antonio Bellido-Jiménez[1], Javier Estévez[1], Amanda P. García-Marín[1]

[1] Projects Engineering Area, Department of Rural Engineering, Civil Constructions and Projects Engineering, University of Córdoba, Spain

Abstract

The presence of missing data in hydrometeorological datasets is a common problem, usually due to sensor malfunction, deficiencies in records storage and transmission, or other recovery procedures issues. These missing values are the primary source of problems when analyzing and modeling their spatial and temporal variability. Thus, accurate gap-filling techniques for rainfall time series are necessary to have complete datasets, which is crucial in studying climate change evolution. In this work, several machine learning models have been assessed to gap-fill rainfall data, using different approaches and locations in the semiarid region of Andalusia (Southern Spain). Based on the obtained results, the use of neighbor data, located within a 50 km radius, highly outperformed the rest of the assessed approaches, with RMSE (root mean squared error) values up to 1.246 mm/day, MBE (mean bias error) values up to −0.001 mm/day, and R^2 values up to 0.898. Besides, inland area results outperformed coastal areas in most locations, arising the efficiency effects based on the distance to the sea (up to an improvement of 63.89% in terms of RMSE). Finally, machine learning (ML) models (especially MLP - Multilayer Perceptron) notably outperformed simple linear regression estimations in the coastal sites, whereas, in inland locations, the improvements were not significant.

1. Introduction

The spatial and temporal analysis of meteorological parameters, such as rainfall is crucial to numerous environmental, hydrological, and agroclimatic studies, as well as optimizing issues, such as water resource management or irrigation scheduling [1–4]. However, one of the most common problems in time series analyses, such as rainfall datasets, is the presence of gaps of different widths, making this task harder to carry out.

This usually results from malfunctioning sensors or data loggers, lack of maintenance, meteorological events, or power outages. Sometimes, the solution is not instantaneous and causes delays because it needs the interaction of qualified personnel. Therefore, before starting with analyses, a common practice is to fill these gaps using different methodologies and applying automatic detection algorithms to detect spurious signals in automated weather stations [5].

Due to its high spatiotemporal variability and the large number of interconnected variables involved, rainfall is one of the most challenging atmospheric variables to characterize, estimate, and forecast [6], especially on a daily basis, with higher volatility and chaotic patterns [7]. A variety of techniques have been developed on both a monthly and daily basis. One of the most frequent algorithms to estimate missing rainfall records is the inverse distance weighting method (IDWM), where the estimated values are calculated with a weighted average (it resorts to the inverse of the distance when assigning the weights) from neighbor stations [8,9]. Another simple method to apply is the gauge mean estimator, which uses an average value of observations from the nearby stations, which can be obtained by optimization, proximity metric, or correlation, among other techniques [10]. Ordinary kriging is a spatially-dependent variance, based on scalar measurements at different locations, where the weights are derived from the distance between the source and the target stations [11–13]. However, these three methods tend to overestimate the number of rainy days and underestimate their magnitudes, and even a negative correlation is found in several reports between close stations [13,14]. Xia et al. [15] evaluated six methodologies (simple arithmetic averaging, inverse distance interpolation, normal ratio method, single best estimator, multiple regression analysis (REG), and closest station method) for estimating missing data in two stations in Germany and concluded that REG consistently obtained the best performance. Teegavarapu and Chandramouli [8] highlighted that the use of the coefficient of correlation provided an improvement in estimating the missing data and recommended the coefficient of correlation weighing method, artificial neural network estimation method, and kriging estimation method for this purpose, due to their conceptually superior performance. Teegavarapu et al. [16] introduced the fixed functional set genetic algorithm method (FFSGAM), eliminating the use of rigid functional forms and weighting approaches for gap-filling. FFSGAM outperformed conventional IDWM. Adhikary et al. [12] developed genetic programming-based ordinary kriging (GPOK) as a new variant of the kriging method, using the genetic programming-derived variogram model and ordinary kriging. GPOK obtained the best results when compared to ANN-based ordinary kriging and traditional ordinary kriging. Different au-thors [17–19] have evaluated the k-nearest-neighbor algorithm, in conjunction with machine learning models, such as multilayer perceptron (MLP), support vector machine (SVM), and random forest (RF), with promising results. Bagirov et al. [20] evaluated cluster-wise linear regression (CLR), using different combinations of maximum and minimum daily air temperature, evaporation, vapor pressure, and solar radiation to predict monthly rainfall in Victoria, Australia. Their results showed a higher performance of CLR against different methods, such as cluster regression-expectation maximization, multiple linear regression, support vector regression (SVR), and MLP. Kajewska-Szkudlarek [21] assessed the use of cluster analysis with SVR to outperform daily rainfall prediction in urban areas.

Additionally, other researchers study the performance of processing algorithms, such as wavelets [22,23], variational mode decomposition (VMD) [24], or singular spectrum analysis (SSA) [25,26]. Estévez et al. [22] evaluated different combinations of wavelet analysis with thermo-pluviometric variables, using MLP in sixteen locations in Spain to forecast monthly rainfall. The results indicated the suitability of the models using thermopluviometric variables without requiring long-term datasets. Partal and Kisis [23] assessed a wavelet analysis, in conjunction with neuro-fuzzy models, to forecast daily rainfall in Turkey. The developed models were significantly superior to traditional machine learning approaches, with a coefficient of determination (R^2) around 0.8–0.9. Li et al. [24] studied the performance of VMD, coupled with an extreme learning machine (ELM) model, to improve monthly rainfall forecasts in the northwest of China. This hybrid model highly outperformed traditional algorithms, with a meager computational cost, due to the non-training requirement of ELM. Filho and Lima [25] evaluated the singular spectrum analysis (SSA) forecasting monthly rainfall in Brazil. Based on the results, it could be concluded that the SSA caterpillar algorithm can deal with the inherent non-stationary nature of rainfall records, extracting its long varying trends and periodic components. Sun et al. [26] assessed SSA in Korea with linear recurrent formulas (LRF) and MLP. MLP obtained the best performance when forecasting monthly rainfall.

Finally, due to the significant advances in computation, deep learning algorithms are gaining very high popularity. In this sense, Kim et al. [27] evaluated the convolutional neural network (CNN), in conjunction with long short-term memory (LSTM), named convLSTM, to nowcast 1 and 2 h in advance, using two years dataset periods. ConvLSTM was able to reduce RMSE by 23% when compared to linear regression. Ha et al. [28] developed a deep belief network model to forecast rainfall one day ahead in Seoul, performing better than MLP. Chen et al. [29] studied the performance of convLSTM with group normalization (GN) to improve the optimization process and employ a multi-sigmoid loss, inspired by the critical success index (CSI) and compared it to the COTREC model. COTREC obtained better performance, in terms of intensity in some areas, whereas con-vLSTM got a generally more reliable forecast.

This study aims to create a daily rainfall estimation model using only precipitation data, with different approaches in semiarid regions, such as Andalusia, to fill possible gaps in precipitation datasets. Additionally, a new approach is tested in daily rainfall estimations, which uses future precipitation values for this purpose. Thus, in this work, several machine learning models (MLP, SVM, and RF) and approaches for estimating missing rainfall data were tested and compared to empirical algorithms, such as linear interpolation (LI), in 14 locations from two different regions of Andalusia (coastal and inland areas) in Southern Spain. The first approach (A) uses neighbor stations' rainfall data of the same gap day and its distance to the target station. All these neighbor stations are located within a radius of up to 50 km, following the recommendations of Barrios et al. [9] on a monthly basis and Estévez et al. [31] on a daily basis. Secondly, a new approach is considered, using only rainfall data (past and future values) from the target station as the model’s inputs. Specifically, two different configurations were tested: (B) one day before and after the gap day and (C) two days before and after the gap day.

The rest of the work is organized as follows. Section 2 shows the information about the locations, the dataset, the theoretical background of the different machine learning (ML) models assessed, the preprocessing algorithms, and the evaluation metrics. Then, in Section 3, the results are reported and discussed. Finally, Section 4 describes the conclusions achieved in this work.

2. Materials and Methods

2.1. Source of Data

This study was carried out in Andalusia, Southern Spain, located in the southwest of Europe. Andalusia is a semiarid region with the following features: the meridians range from 1 to 7° W, the parallels from 37° to 39° N, an elevation above mean sea level from 26 to 822 m, and a total area of 87 268 m2.

The used datasets belong to the Agroclimatic Information Network of Andalusia (RIAA) and can be downloaded at the following link: https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web (accessed on September 8th, 2021). A total of 14 stations, divided into two areas (coastal and inland locations), were evaluated. The first group of areas included Jaen, La Higuera de Arjona, Lina-res, Mancha Real, Marmolejo, Sabiote, and Torreblascopedro, and the second group included Málaga, Antequera, Archidona, Cártama, Churriana, Pizarra, and Vélez. Figure 5. 1 shows their geographical locations, and Table 5. 1 shows their geo-climatic characteristics.

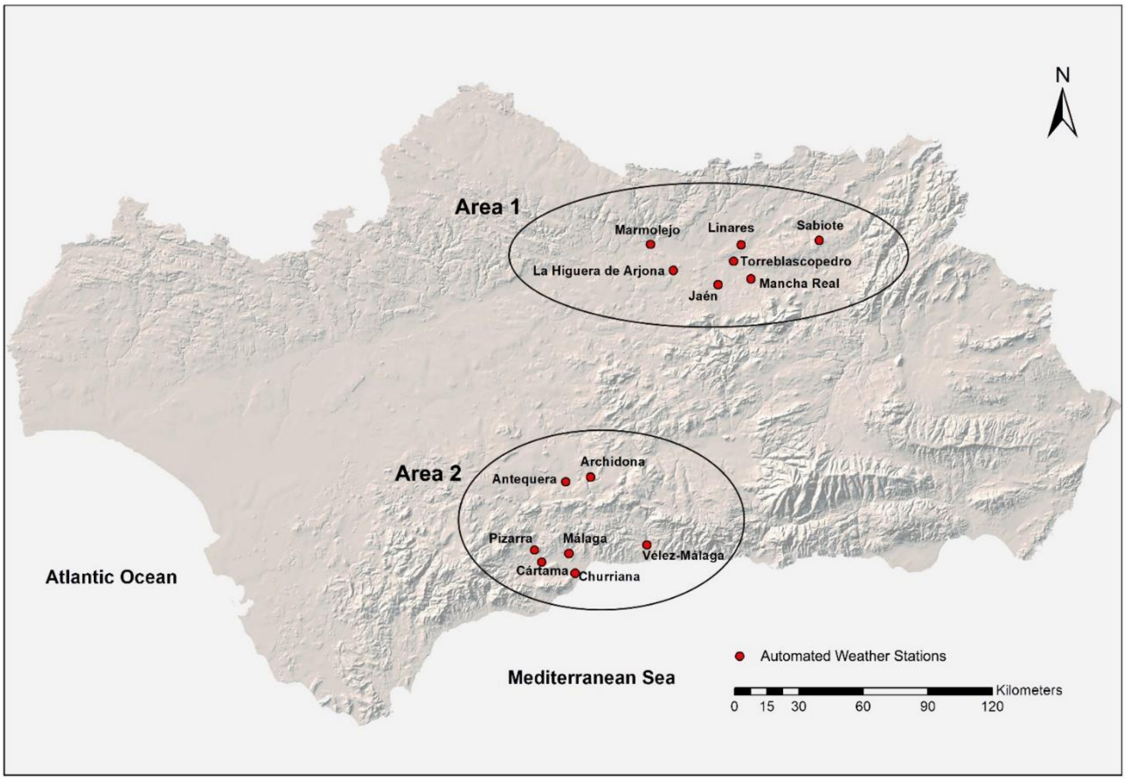


Figure 5. 1. Spatial distribution of the fourteen automated weather stations used in this work.

Table 5. 1. Geo-climatic characteristics of the AWS assessed in this work (lat.: latitude; long.: longitude; alt.: elevation above mean sea level).

Station	**Alt. [m]**	**Lat. [ºN]**	**Long. [ºW]**	**Mean Annual Rainfall [mm]**	**Time-Period (Number of Days)**
Area 1:					
Jaen (JAE)	299	37.89	3.77	446.54	From April 2001 to June 2021 (7361)
La Higuera de Arjona (ARJ)	257	37.95	4.00	477.68	From January 2001 to June 2021 (7456)
Linares (LIN)	432	38.07	3.65	466.70	From August 2000 to June 2021 (7601)
Mancha Real (MAN)	407	37.92	3.60	390.86	From August 2000 to June 2021 (7602)
Marmolejo (MAR)	208	38.06	4.13	523.36	From September 2000 to June 2021 (7590)
Sabiote (SAB)	791	38.08	3.24	446.98	From August 2000 to June 2021 (7615)
TorreblascoPedro (TOR)	275	37.99	3.69	434.37	From August 2000 to June 2021(7615)

Area 2:					
Antequera (ANT)	440	37.03	4.56	444.72	From November 2000 to June 2021 (7512)
Archidona (ARC)	516	37.08	4.43	457.83	From December 2000 to June 2021 (7483)
Cártama (CAR)	78	36.72	4.68	490.51	From June 2001 to June 2021 (7300)
Churriana (CHU)	17	36.67	4.50	510.32	From February 2001 to June 2021 (7426)
Málaga (MAL)	55	36.76	4.54	461.63	From October 2000 to June 2021 (7546)
Pizarra (PIZ)	71	36.77	4.72	463.47	From January 2001 to June 2021 (7447)
Vélez (VEL)	33	36.80	4.13	490.49	From October 2000 to June 2021 (7546)

2.2. Methodology

An essential prerequisite to guarantee reliable results using raw meteorological data is the application of quality assurance procedures. The quality control guidelines, reported by Estévez et al. [31], were followed, as well as the procedure to detect spurious precipitation signals from automated weather stations (AWS), also Estévez et al. [5].

Afterward, data preprocessing was required for every approach, obtaining the corresponding input configuration, according to every strategy (see Table 5. 2). Three different methodologies were evaluated: approach (i)—the use of rainfall neighbor data and its distance data to estimate the precipitation values at a different site (all locations are located within a 50 km radius); approach (ii)—the use of one day before and ahead rainfall da-ta values from the target station; and approach (iii)—the use of two days before and ahead rainfall data values from the target station.

Later, in order to tune all the different hyperparameters from the different models, train them, and evaluate their performance, the full dataset was split into training, validation, and testing. The train (to fit all the final weights and biases from the final model) and test dataset (never-seen data to assess the performance) were randomly split into 70% and 30%, respectively. Prior to this stage, it is necessary to determine all the hyperparameters of the models (such as the number of hidden layers and neurons in a multilayer perceptron). For this purpose, the training dataset was divided into train_2 and validation (ran-dom 80% and 20%, respectively) to train and test the different hyperparameters until the fittest set is found. It is worth noting that the seed used in the random algorithm is the same in all cases, so all assessed models (from different approaches) have the same train, test, and validation dataset. Then, the Bayesian optimization algorithm took place, where different hyperparameters were tested, using the validation dataset, until the fittest set was found. Afterward, the entire train dataset from the initial split was used to adjust all the different weights and biases. Finally, the performance accuracy was assessed, using the testing dataset, which was never seen during previous processes. All this methodology is shown in a flowchart in Figure 5. 2.

Table 5. 2. . Inputs configurations of the different models and approaches assessed. DOY represents the day of year, P corresponds to precipitation, D corresponds to distance, and i represents an index to the dataset-specific day.

Target Station	Inputs Approach A	Inputs Approach B	Inputs Approach C
Area 1:			
Jaen	$DOY(i) + P_{ARJ}(i) + D_{JAE\text{-}LIN} + P_{LIN}(i) + D_{JAE\text{-}LIN} + P_{MAN}(i) + D_{JAE\text{-}MAN} + P_{MAR}(i) + D_{JAE\text{-}MAR} + P_{SAB}(i) + D_{JAE\text{-}SAB} + P_{TOR}(i) + D_{JAE\text{-}TOR}$	$DOY(i) + P_{JAE}(i-1) + P_{JAE}(i+1)$	$DOY(i) + P_{JAE}(i-1) + P_{JAE}(i-2) + P_{JAE}(i+1) + P_{JAE}(i+2)$
La Higuera de Arjona	$DOY(i) + P_{JAE}(i) + D_{ARJ\text{-}JAE} + P_{LIN}(i) + D_{ARJ\text{-}LIN} + P_{MAN}(i) + D_{ARJ\text{-}MAN} + P_{MAR}(i) + D_{ARJ\text{-}MAR} + P_{SAB}(i) + D_{ARJ\text{-}SAB} + P_{TOR}(i) + D_{ARJ\text{-}TOR}$	$DOY(i) + P_{ARJ}(i-1) + P_{ARJ}(i+1)$	$DOY\ (i) + P_{ARJ}(i-1) + P_{ARJ}(i-2) + P_{ARJ}(i+1) + P_{ARJ}(i+2)$
Linares	$DOY(i) + P_{JAE}(i) + D_{LIN\text{-}JAE} + P_{ARJ}(i) + D_{LIN\text{-}ARJ} + P_{MAN}(i) + D_{LIN\text{-}MAN} + P_{MAR}(i) + D_{LIN\text{-}MAR} + P_{SAB}(i) + D_{LIN\text{-}SAB} + P_{TOR}(i) + D_{LIN\text{-}TOR}$	$DOY(i) + P_{LIN}(i-1) + P_{LIN}(i+1)$	$DOY\ (i) + P_{LIN}(i-1) + P_{LIN}(i-2) + P_{LIN}(i+1) + P_{LIN}(i+2)$
Mancha Real	$DOY(i) + P_{JAE}(i) + D_{MAN\text{-}JAE} + P_{ARJ}(i) + D_{MAN\text{-}ARJ} + P_{LIN}(i) + D_{MAN\text{-}LIN} + P_{MAR}(i) + D_{MAN\text{-}MAR} + P_{SAB}(i) + D_{MAN\text{-}SAB} + P_{TOR}(i) + D_{MAN\text{-}TOR}$	$DOY(i) + P_{MAN}(i-1) + P_{MAN}(i+1)$	$DOY\ (i) + P_{MAN}(i-1) + P_{MAN}(i-2) + P_{MAN}(i+1) + P_{MAN}(i+2)$
Marmolejo	$DOY(i) + P_{JAE}(i) + D_{MAR\text{-}JAE} + P_{ARJ}(i) + D_{MAR\text{-}ARJ} + P_{LIN}(i) + D_{MAR\text{-}LIN} + P_{MAN}(i) + D_{MAR\text{-}MAN} + P_{SAB}(i) + D_{MAR\text{-}SAB} + P_{TOR}(i) + D_{MAR\text{-}TOR}$	$DOY(i) + P_{MAR}(i-1) + P_{MAR}(i+1)$	$DOY\ (i) + P_{MAR}(i-1) + P_{MAR}(i-2) + P_{MAR}(i+1) + P_{MAR}(i+2)$
Sabiote	$DOY(i) + P_{JAE}(i) + D_{SAB\text{-}JAE} + P_{ARJ}(i) + D_{SAB\text{-}ARJ} + P_{LIN}(i) + D_{SAB\text{-}LIN} + P_{MAN}(i) + D_{SAB\text{-}MAN} + P_{MAR}(i) + D_{SAB\text{-}MAR} + P_{TOR}(i) + D_{SAB\text{-}TOR}$	$DOY(i) + P_{SAB}(i-1) + P_{SAB}(i+1)$	$DOY\ (i) + P_{SAB}(i-1) + P_{SAB}(i-2) + P_{SAB}(i+1) + P_{SAB}(i+2)$
TorreblascoPedro	$DOY(i) + P_{JAE}(i) + D_{TOR\text{-}JAE} + P_{ARJ}(i) + D_{TOR\text{-}ARJ} + P_{LIN}(i) + D_{TOR\text{-}LIN} + P_{MAN}(i) + D_{TOR\text{-}MAN} + P_{MAR}(i) + D_{TOR\text{-}MAR} + P_{SAB}(i) + D_{TOR\text{-}SAB}$	$DOY(i) + P_{TOR}(i-1) + P_{TOR}(i+1)$	$DOY\ (i) + P_{TOR}(i-1) + P_{TOR}(i-2) + P_{TOR}(i+1) + P_{TOR}(i+2)$
Area 2:			

Antequera	$DOY(i) + P_{ARC}(i) + D_{ANT-ARC} + P_{CAR}(i) + D_{ANT-CAR} + P_{CHU}(i) + D_{ANT-CHU} + P_{MAL}(i) + D_{ANT-MAL} + P_{PIZ}(i) + D_{ANT-PIZ} + P_{VEL}(i) + D_{ANT-VEL}$	$DOY(i) + P_{ANT}(i-1) + P_{ANT}(i+1)$	$DOY (i) + P_{ANT}(i-1) + P_{ANT}(i-2) + P_{ANT}(i+1) + P_{ANT}(i+2)$
Archidona	$DOY(i) + P_{ANT}(i) + D_{ARC-ANT} + P_{CAR}(i) + D_{ARC-CAR} + P_{CHU}(i) + D_{ARC-CHU} + P_{MAL}(i) + D_{ARC-MAL} + P_{PIZ}(i) + D_{ARC-PIZ} + P_{VEL}(i) + D_{ARC-VEL}$	$DOY(i) + P_{ARC}(i-1) + P_{ARC}(i+1)$	$DOY (i) + P_{ARC}(i-1) + P_{ARC}(i-2) + P_{ARC}(i+1) + P_{ARC}(i+2)$
Cártama	$DOY(i) + P_{ANT}(i) + D_{CAR-ANT} + P_{ARC}(i) + D_{CAR-ARC} + P_{CHU}(i) + D_{CAR-CHU} + P_{MAL}(i) + D_{CAR-MAL} + P_{PIZ}(i) + D_{CAR-PIZ} + P_{VEL}(i) + D_{CAR-VEL}$	$DOY(i) + P_{CAR}(i-1) + P_{CAR}(i+1)$	$DOY (i) + P_{CAR}(i-1) + P_{CAR}(i-2) + P_{CAR}(i+1) + P_{CAR}(i+2)$
Churriana	$DOY(i) + P_{ANT}(i) + D_{CHU-ANT} + P_{ARC}(i) + D_{CHU-ARC} + P_{CAR}(i) + D_{CHU-CAR} + P_{MAL}(i) + D_{CHU-MAL} + P_{PIZ}(i) + D_{CHU-PIZ} + P_{VEL}(i) + D_{CHU-VEL}$	$DOY(i) + P_{CHU}(i-1) + P_{CHU}(i+1)$	$DOY (i) + P_{CHU}(i-1) + P_{CHU}(i-2) + P_{CHU}(i+1) + P_{CHU}(i+2)$
Málaga	$DOY(i) + P_{ANT}(i) + D_{MAL-ANT} + P_{ARC}(i) + D_{MAL-ARC} + P_{CAR}(i) + D_{MAL-CAR} + P_{CHU}(i) + D_{MAL-CHU} + P_{PIZ}(i) + D_{MAL-PIZ} + P_{VEL}(i) + D_{MAL-VEL}$	$DOY(i) + P_{MAL}(i-1) + P_{MAL}(i+1)$	$DOY (i) + P_{MAL}(i-1) + P_{MAL}(i-2) + P_{MAL}(i+1) + P_{MAL}(i+2)$
Pizarra	$DOY(i) + P_{ANT}(i) + D_{PIZ-ANT} + P_{ARC}(i) + D_{PIZ-ARC} + P_{CAR}(i) + D_{PIZ-CAR} + P_{CHU}(i) + D_{PIZ-CHU} + P_{MAL}(i) + D_{PIZ-MAL} + P_{VEL}(i) + D_{PIZ-VEL}$	$DOY(i) + P_{PIZ}(i-1) + P_{PIZ}(i+1)$	$DOY (i) + P_{PIZ}(i-1) + P_{PIZ}(i-2) + P_{PIZ}(i+1) + P_{PIZ}(i+2)$
Vélez	$DOY(i) + P_{ANT}(i) + D_{VE-ANT} + P_{ARC}(i) + D_{VEL-ARC} + P_{CAR}(i) + D_{VEL-CAR} + P_{CHU}(i) + D_{VEL-CHU} + P_{MAL}(i) + D_{VEL-MAL} + P_{PIZ}(i) + D_{VEL-PIZ}$	$DOY(i) + P_{VEL}(i-1) + P_{VEL}(i+1)$	$DOY (i) + P_{VEL}(i-1) + P_{VEL}(i-2) + P_{VEL}(i+1) + P_{VEL}(i+2)$

Besides, after splitting the dataset into train and test, a standardization was carried out, which is highly recommended to outperform machine learning models, especially neural network-based models [32]. This can be expressed as Eq. 5.1:

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{Eq. 5.1}$$

where x represents the input data and x¯and σ correspond to the mean and standard deviation of the training dataset, respectively, and x^* is the standardized data.

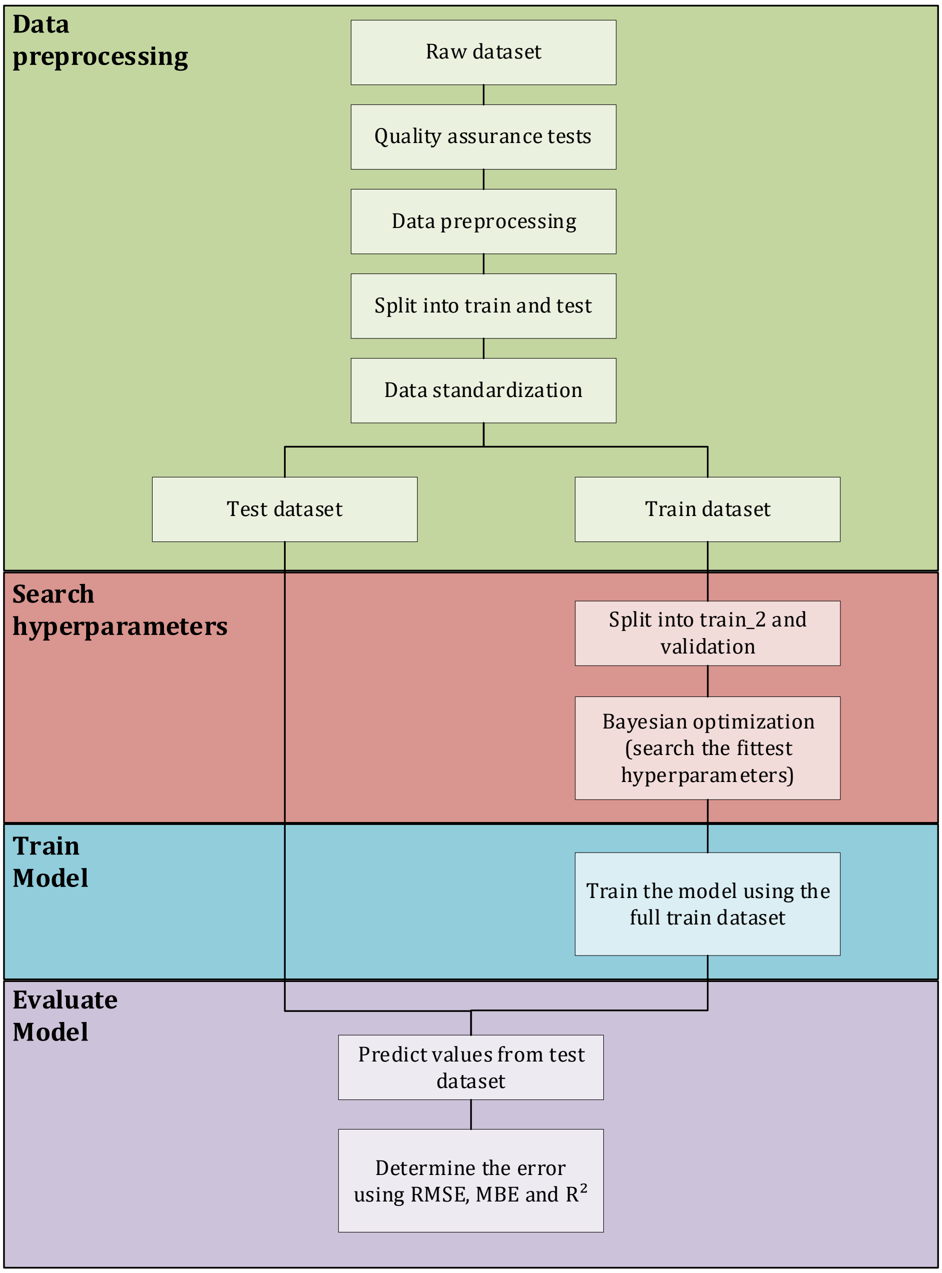


Figure 5. 2. Methodology flowchart.

2.3. Multilayer Perceptron (MLP)

Multilayer perceptron is one of the most used models in different sectors, especially in hydrology [22,33]. Its functionality is based on neurons in the biological nervous system, where many interconnected neurons work together to generate an interaction, based on different stimuli. It is structured in three types of layers, the input and output correspond to the input and output of the model, respectively, as well as the hidden layer, where neurons are located. The activation function determines the output of a node, given a set of inputs. For example, rectified linear output (ReLU) represents a ramp for positive input values. The process in which the neurons learn (value adjustment of weights and biases) is carried out automatically, which is why this layer is called hidden. ADAM, a very common algorithm for this purpose, uses squared gradients to scale the learning rate and a moving average of the gradients.

A single neuron mathematical logic is represented in Figure 5. 3, where w represents the weight and b is the bias factor.

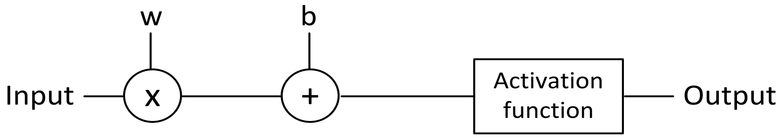
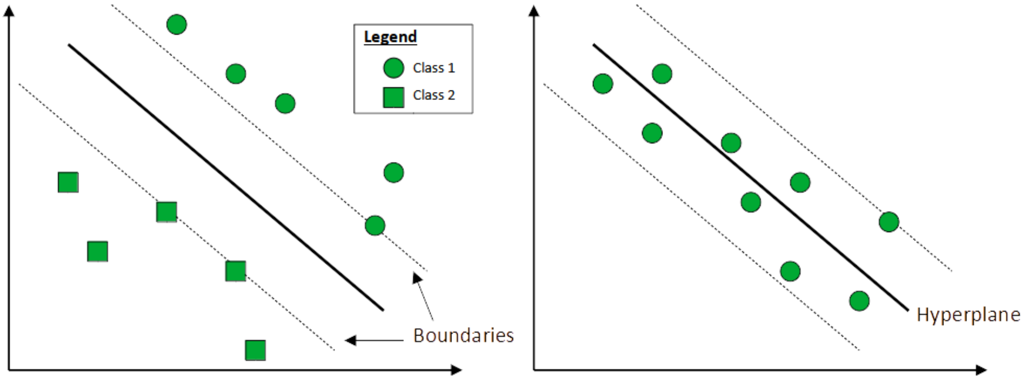


Figure 5. 3. One neuron control logic

For further information, the following works can be reviewed [34,35]

2.4. Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised machine learning model that analyzes data for classification and regression tasks (also known as support vector regression (SVR)). For classification tasks, its functionality is based on searching the fittest hyperplane to separate different data points' classes (classification). On regression, it finds the hyperplane and margins that fit all of them (regression). Thus, an easy way to understand SVM for regression is similar to a linear regression, where a hyperplane (that includes the data) is searched while having the flexibility to define how much error is considered acceptable. Figure 5. 4 shows an example of SVM for classification (a) and regression (b).



(a) **(b)**

Figure 5. 4. Support Vector Machine for classification (a) and regression (b).

The main feature of SVM models is the use of kernels (linear, sigmoid, or gaussian, among others) to enable operation in a high-dimensional feature map, where the number of features is greater than the number of observations.

SVM models are often used in rainfall forecasts, with promising results [36–38]. For further details, the following work can be reviewed [36,37].

2.5. Random Forest (RF)

Random Forest (RF) was first introduced by [39] as a supervised learning algorithm, where the "term" forest defines that it is built as an ensemble of decision tree models. The general idea is that the conjunction of multiple models increases the overall result. Additionally, RF introduces an extra-randomness when the number of trees starts to grow.

Instead of searching for the best feature when splitting nodes, it searches for the best features among a random subset of them. The maximum number of features can be defined in sci-kit-learn as auto, sqrt, log2, none, or the exact number of maximum features (where au-to and sqrt refer to the squared root of the initial number of features, log2 refers to the logarithm base 2 of the number of features, and none is to use all the features). This results in a broader diversity and, as a consequence, a better final performance.

Other researchers have already assessed RF in rainfall with promising results [40–42]. For further details, the following work can be revised [42].

2.6. Bayesian Optimization

One of the critical aspects of machine learning models' efficiency is hyperparameter selection. Depending on whether the correct values have been set, the performance can dramatically change from outstanding to very poor results. A common practice in the scientific community uses a trial-and-error technique [22], where different values are evaluated, varying from dozens to thousands of possibilities. However, this method is far from efficient because if the hyperparameter space is ample, the algorithm (apart from being very slow) wastes significant time in non-promising configurations. On the other hand, when the hyperparameter space is tiny, an accurate hyperparameter configuration set may be missing, despite being quick. To solve this problem, several algorithms have been assessed in different works. In [43], the authors studied the effectiveness of particle swarm optimization (PSO) and genetic algorithm (GA) to predict the monthly rainfall with MLP in a subtropical monsoon climate in Guilin, China. Wang et al. [44] assessed an artificial bee colony (ABC) with MLP to forecast rainfall values in 17 stations in the Wujiang River Basin. Banadkooki et al. [35] evaluated the flow regime optimization algorithm (FRA) with MLP and SVM to forecast monthly rainfall values in Iran.

In this study, Bayesian optimization was used, due to its high popularity in new automated machine learning (AML) models [45–48] and its good performance in [34,49]. It was first introduced by Wang et al. [44] as an algorithm, based on the Bayes theorem, to search the minimum/maximum function. Part of its popularity is due to its close relation to human behavior when tuning hyperparameters [50,51]. The prior results are taken into account to choose the following promising values to test, following the next four-step procedure: (1) the hyperparameter space is defined, which limits the values of the hyperparameter space; (2) the algorithm considers previous evaluations, in order to choose the following set of values to be assessed (acquisition function)—two kinds of possibilities can be handled, exploitation (consists of testing hyperparameters values that are assumed to be optimal) and exploration (the opposite of exploitation, to identify new best options); (3) to assess the new hyperparameter configuration using an objective function; and (4) if the optimization process has not finished yet, it goes to the second point. In this work, this algorithm was implemented using Python and the Scikit-Optimize library, following the instructions of Bellido-Jiménez et al. [34]. All the final hyperparameter sets, used for each model, approach, and location, can be seen in Table 3.

Table 5. 3. Hyperparameter set for each model, approach, and location, after carrying out Bayesian optimi-zation, where activation represents the activation function, the optimizer represents the optimizer function, epochs represents the number of epochs, neurons

			Approaches:		
Location	**Models**	**Hyperparameters**	**A**	**B**	**C**
La Higuera de Arjona	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	87	53
		neurons	(20, 20)	(9, 15, 10)	(6, 15, 9)
	SVM	kernel	RBF	RBF	poly
		c	10.0	10.0	1.855
		epsilon	0.01	0.01	0.01
	RF	n_estimators	100	100	91
		max_features	sqrt	auto	log2
Jaen	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	92	61	98
		neurons	(20, 20)	(2, 1, 12)	(1, 10, 8)
	SVM	kernel	linear	linear	RBF
		c	1.758	9.730	10.0
		epsilon	0.739	0.01	0.01
	RF	n_estimators	94	95	100
		max_features	auto	log2	log2
Linares	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	100	100
		neurons	(20, 20)	(1, 1, 1)	(1, 1, 1)
	SVM	kernel	linear	RBF	RBF
		c	10.0	4.023	10.0
		epsilon	0.01	0.018	0.01
	RF	n_estimators	100	97	80
		max_features	auto	sqrt	log2
Mancha Real	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	99	100
		neurons	(20, 20)	(5, 14)	(1, 1, 1)
	SVM	kernel	RBF	RBF	RBF
		c	10.0	6.235	9.211
		epsilon	0.01	0.010	0.01
	RF	n_estimators	75	41	46
		max_features	auto	sqrt	log2
Marmolejo	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	96	10
		neurons	(20, 6)	(5, 3, 11)	(1, 11)
	SVM	kernel	linear	RBF	RBF
		c	10.0	4.350	9.970
		epsilon	0.01	0.01	0.01
	RF	n_estimators	100	31	100
		max_features	auto	auto	sqrt
Sabiote	MLP	activation	ReLU	ReLU	ReLU

		optimizer	ADAM	ADAM	ADAM
		epochs	100	100	95
		neurons	(20, 20)	(1, 1, 1)	(2, 11, 9)
	SVM	kernel	linear	RBF	RBF
		c	10.0	10.0	10.0
		epsilon	0.01	0.01	0.01
	RF	n_estimators	72	39	57
		max_features	log2	log2	log2
Torreblascopedro	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	72	73
		neurons	(20, 12)	(1, 4, 13)	(5, 1, 17)
	SVM	kernel	linear	RBF	poly
		c	3.795	3.108	6.205
		epsilon	0.01	0.01	0.012
	RF	n_estimators	81	64	94
		max_features	log2	sqrt	sqrt
Antequera	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	200	174	61
		neurons	(13, 8)	(5, 2, 20)	(14, 11, 13)
	SVM	kernel	linear	RBF	RBF
		c	8.684	7.627	4.981
		epsilon	0.225	0.01	0.014
	RF	n_estimators	55	94	41
		max_features	auto	auto	log2
Archidona	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	94	40	11
		neurons	(13, 5, 18)	(11, 12, 1)	(16, 11, 19)
	SVM	kernel	linear	poly	RBF
		c	7.246	4.531	4.104
		epsilon	0.01	0.01	0.013
	RF	n_estimators	81	93	100
		max_features	auto	auto	sqrt
Cártama	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	129	10	112
		neurons	(8, 13, 17)	(1, 1, 1)	(14, 17, 6)
	SVM	kernel	linear	RBF	poly
		c	6.273	7.830	3.862
		epsilon	0.01	0.01	0.01
	RF	n_estimators	92	10	38
		max_features	auto	sqrt	log2
Churriana	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	180	200	70
		neurons	(20, 20, 18)	(1, 1, 1)	(5, 15, 7)
	SVM	kernel	linear	RBF	RBF
		c	10.0	10.0	5.963
		epsilon	0.01	0.01	0.01
	RF	n_estimators	100	40	36

		max_features	log2	log2	sqrt
Málaga	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	158	97	127
		neurons	(20, 20, 20)	(17, 11, 10)	(13, 4, 16)
	SVM	kernel	linear	RBF	RBF
		c	8.784	9.999	6.952
		epsilon	0.01	0.01	0.011
	RF	n_estimators	69	14	10
		max_features	log2	log2	log2
Pizarra	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	192	94	171
		neurons	(13, 15, 8)	(3, 4, 6)	(14, 1, 6)
	SVM	kernel	linear	RBF	RBF
		c	7.642	10.0	4.031
		epsilon	0.015	0.01	0.01
	RF	n_estimators	76	45	95
		max_features	auto	sqrt	sqrt
Vélez-Málaga	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	200	180	139
		neurons	(20, 20, 20)	(15, 13, 10)	(8, 2, 10)
	SVM	kernel	RBF	RBF	RBF
		c	10.0	6.032	10.0
		epsilon	0.01	0.01	0.01
	RF	n_estimators	72	62	78
			sqrt	log2	sqrt

2.7. Evaluation Metrics

To assess the efficiency of the developed models, the statistics root mean square error (RMSE), mean bias error (MBE), and coefficient of determination (R^2) were used. All of them are mathematically expressed as Equations 5.2 to 5.4, respectively:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2} \qquad \text{Eq. 5.2}$$

$$MBE = \frac{1}{m}\sum_{i=1}^{m} meas_i - pred_i \qquad \text{Eq. 5.3}$$

$$R^2 = \frac{\left(\sum_{i=1}^{m}(meas_i - \overline{meas})(pred_i - \overline{pred})\right)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2 \sum_{i=1}^{m}\left(pred_i - \overline{pred}\right)^2} \qquad \text{Eq. 5.4}$$

where n represents the number of prediction days, $meas_i$ corresponds to the measured value for a specific day, $pred_i$ is the predicted value, i represents every single gap day, and μ corresponds to the mean.

3. Results and Discussion

In order to help the reproducibility of this work, the best ML models were uploaded to an open-access repository in Github (https://github.com/Smarity/gap-filling-precipitation-atmosphere-special-issue.git - accessed on September 8th, 2021).

3.1. Using Neighbor Stations

Table 5. 4 shows the RMSE, MBE, and R^2 performances for all locations in Area 1 (inland locations) using the first approach (A), information from other AWS located within 50 km. In Higuera de Arjona, MLP obtained the best RMSE and R^2 values (RMSE = 1.363 mm/day and R^2 = 0.894), very close to RF (RMSE = 1.384 mm/day and R^2 = 0.889). In terms of MBE, LI outperformed the rest of the ML models (MBE = −0.008 mm/day), followed closely by MLP and RF (MBE = 0.016 mm/day and MBE = 0.026 mm/day, respectively). In Jaen, all ML models outperformed LI in RMSE and R^2, where MLP obtained the best values (RMSE = 1.767 mm/day and R^2 = 0.827), whereas RF beat the rest, regarding MBE (MBE = 0.023 mm/day). In Linares, RF and LI obtained the best performance, in terms of MBE (MBE = 0.001 mm/day and MBE = −0.001 mm/day). Moreover, MLP outperformed the others, regarding RMSE and R^2 (RMSE = 1.723 mm/day and R^2 = 0.817), followed closely by RF (RMSE = 1.730 mm/day and R^2 = 0.815). In Mancha Real, MLP outperformed the other models in all statistics (RMSE = 1.662 mm/day, MBE = −0.072 mm/day, and R^2 = 0.831), whereas SVM was the worst (RMSE = 1.948 mm/day, MBE = −0.195 mm/day, and R^2 = 0.780). In Marmolejo, with the highest mean annual rainfall (523.36 mm/year), the performance, in terms of RMSE and R^2, showed that RF obtained the best values (RMSE = 2.129 mm/day and R^2 = 0.801), followed closely by SVM (RMSE = 2.154 mm/day and R^2 = 0.795) and MLP (RMSE = 2.176 mm/day and R^2 = 0.791). In Sabiote, the location with the highest altitude, MLP obtained the best performance in RMSE and R^2 (RMSE = 2.049 mm/day and R^2 = 0.752), but LI beat ML in MBE (MBE = −0.006 mm/day). Finally, in Torreblascopedro, SVM outperformed the rest for all statistics (RMSE = 1.246 mm/day, MBE = −0.005, and R^2 = 0.894), being the most accurate from this first region. It is worth noting that MLP generally obtained the best results, regarding RMSE and R^2, in most locations, whereas RF and LI obtained the best values for MBE. Additionally, even though ML outperformed LI in all locations, the average improvement was not very significant.

Table 5. 5 shows the RMSE, MBE, and R^2 values for all locations and models in the coastal locations (Area 2). In Antequera, MLP beat the other models for all statistics (RMSE = 1.596 mm/day, MBE = 0.035 mm/day, and R^2 = 0.875), sharing the same R^2 performance with SVM (R^2 = 0.875). All ML models highly outperformed LI, considering all statistics (especially RMSE and R^2), except for MBE using SVM. In Archidona, MLP also obtained the most accurate modeling in RMSE and R^2 (RMSE = 1.811 mm/day and R^2 = 0.844), followed closely to SVM (RMSE = 1.817 mm/day and R^2 = 0.844). Regarding MBE, RF outperformed the rest (MBE = −0.019 mm/day). In Cártama, RF obtained the best MBE value (MBE = 0.002 mm/day), whereas SVM got the best RMSE and R^2 performance (RMSE = 2.502 mm/day and R^2 = 0.778). In Churriana, MLP highly outperformed the rest, in terms of RMSE and R^2 (RMSE = 2.192 mm/day and R^2 = 0.876), whereas RF beat MLP in MBE (MBE = 0.019 mm/day and MBE = −0.052 mm/day,

respectively). In Málaga, RF obtained the best values for RMSE and MBE (RMSE = 2.433 mm/day and MBE = 0.012 mm/day), whereas MLP got the most accurate values for R^2 ($R^2 = 0.830$). In Pizarra, all models obtained very similar performance (even LI). RMSE ranged from 2.032 mm/day (by MLP and SVM) to 2.108 mm/day (by LI), MBE ranged from 0.039 mm/day (by RF) to −0.112 mm/day (by SVM), and R^2 ranged from 0.842 (by LI) to 0.854 (by MLP). Finally, in Vélez, MLP outperformed the rest of the models, in terms of RMSE and R^2 (RMSE = 3.219 mm/day and $R^2 = 0.742$), while RF obtained the best MBE performance (MBE = −0.020 mm/day), followed closely to MLP (MBE = −0.074 mm/day). Generally, the results obtained by ML highly outperformed LI in most locations and statistics, except for MBE, in which LI obtained very accurate results. Thus, the use of ML models to gap-fill daily rainfall data is highly recommended for coastal sites, performing significantly better than LI, arising the effect of sea distance in rainfall modeling. Eventually, in Figure 5. 5, all these RMSE, MBE, and R^2 values, from both areas and all models, are represented in a scatter plot. Due to the different performances between the ML models, it can be stated that MLP obtained the best results, or very close to them, in most locations. On the other hand, SVM had accurate performances in coastal sites, whereas the behavior was not so good in inland locations. Finally, RF behaved opposite to SVM, having an accurate performance on inland locations and a worse modeling on inland sites.

Table 5. 4. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the first area (inland locations), using data from neighbor stations. The best values for each site are in bold

Stations (Area 1)	**Model**	**RMSE [mm/day]**	**MBE [mm/day]**	**R^2**
La Higuera de Arjona	MLP	**1.363**	0.016	**0.894**
	SVM	1.800	−0.106	0.818
	RF	1.384	0.026	0.889
	LI	1.502	**−0.008**	0.869
Jaen	MLP	**1.767**	−0.097	**0.827**
	SVM	1.822	−0.064	0.823
	RF	1.880	**0.023**	0.804
	LI	1.916	0.051	0.797
Linares	MLP	**1.723**	0.083	**0.817**
	SVM	1.808	−0.106	0.798
	RF	1.730	**0.001**	0.815
	LI	1.896	−**0.001**	0.784
Mancha Real	MLP	**1.662**	**−0.072**	**0.831**
	SVM	1.948	−0.195	0.780
	RF	1.730	−0.078	0.816
	LI	1.852	0.110	0.790
Marmolejo	MLP	2.176	−0.187	0.791
	SVM	2.154	−0.169	0.795
	RF	**2.129**	**0.041**	**0.801**

	LI	2.392	−0.249	0.753
Sabiote	MLP	**2.049**	−0.101	**0.752**
	SVM	2.135	−0.224	0.739
	RF	2.105	−0.061	0.740
	LI	2.112	**−0.006**	0.742
Torreblascopedro	MLP	1.270	−0.035	0.894
	SVM	**1.246**	**−0.005**	**0.898**
	RF	1.359	0.019	0.878
	LI	1.277	0.047	0.894
Mean values		1.792	−0.048	0.815

Table 5. 5. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the second area (coastal locations), using data from neighbor stations. The best values for each site are in bold.

Stations (Area 2)	**Model**	**RMSE [mm/day]**	**MBE [mm/day]**	**R^2**
Antequera	MLP	**1.595**	**0.035**	**0.875**
	SVM	1.632	−0.104	**0.875**
	RF	2.009	0.042	0.799
	LI	2.839	0.100	0.684
Archidona	MLP	**1.811**	−0.043	**0.844**
	SVM	1.817	−0.168	**0.844**
	RF	2.002	**−0.019**	0.809
	LI	3.286	−0.041	0.594
Cártama	MLP	2.640	−0.075	0.756
	SVM	**2.502**	−0.106	**0.778**
	RF	2.820	**0.002**	0.737
	LI	2.630	0.061	0.756
Churriana	MLP	**2.192**	−0.052	**0.876**
	SVM	2.465	−0.147	0.860
	RF	2.315	**0.019**	0.862
	LI	2.973	−0.061	0.790
Málaga	MLP	2.485	0.099	**0.830**
	SVM	2.448	−0.170	0.825
	RF	**2.433**	**0.012**	0.816
	LI	2.610	0.04	0.785
Pizarra	MLP	**2.032**	0.043	**0.854**
	SVM	2.083	−0.112	0.853
	RF	**2.032**	**0.039**	**0.854**
	LI	2.108	0.079	0.842
Vélez-Málaga	MLP	**3.219**	−0.074	**0.742**
	SVM	3.531	−0.376	0.706
	RF	3.306	**−0.020**	0.719

	LI	3.489	−0.157	0.692
Mean values		2.475	−0.041	0.794

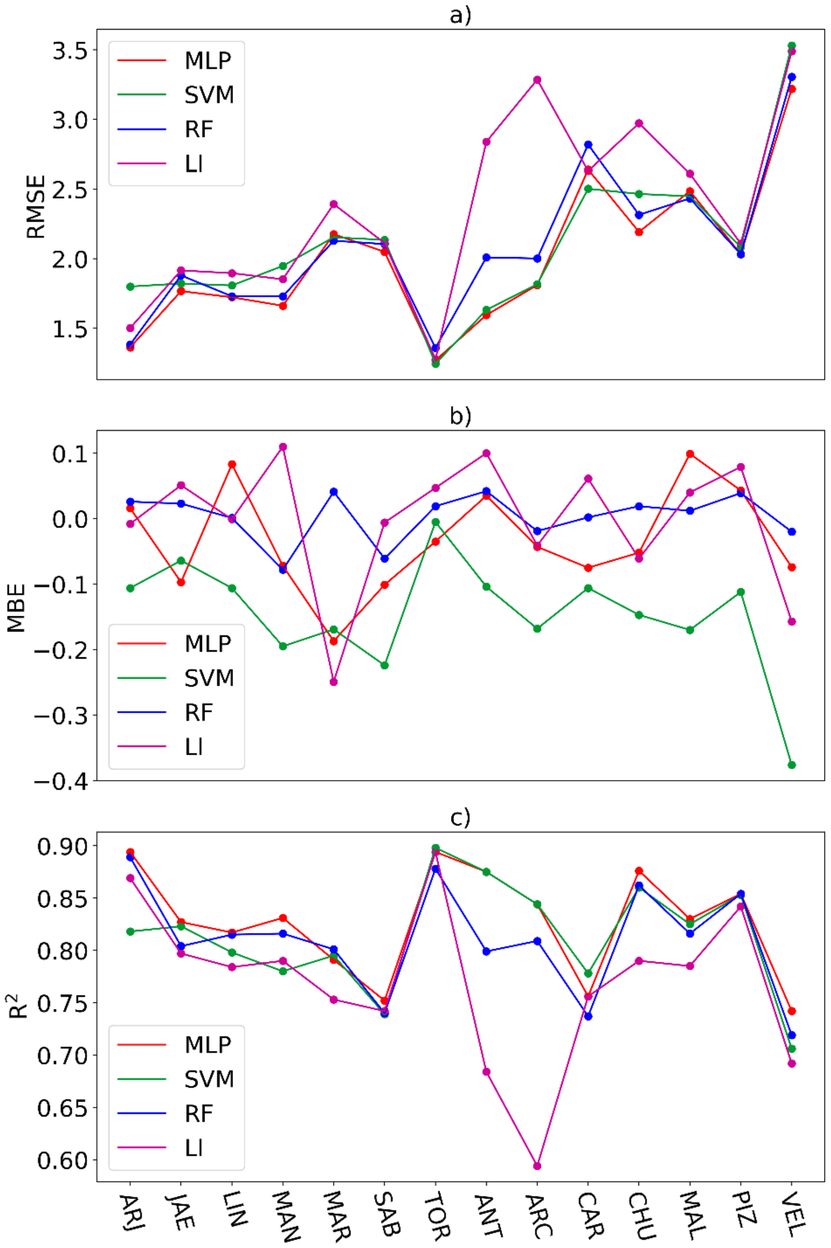


Figure 5. 5. RMSE (a), MBE (b), and R^2 (c) values from testing dataset for all stations and models (MLP, SVM, RF, and LI), using only precipitation data from neighbor stations.

3.2. Using Data from the Target Station

Tables 5. 6 and 5. 7 show the RMSE and MBE values for the inland and coastal locations, using two different approaches, one day after and before (approach B) and two days after and before (approach C), as inputs, respectively. Generally, all the results are much worse than in Tables 2 and 3, for all cases. Mancha Real obtained an RMSE value above 4.0 for all models, whereas Churriana got the worse values (RMSE > 6.0 mm/day). In terms of MBE, La Higuera de Arjona obtained the best performance (MBE = −0.021

mm/day) using MLP and approach B, whereas Marmolejo was the worst (MBE = −1.459 mm/day), using MLP, and this same approach. Finally, in terms of R^2, the values obtained are low, from $R^2 = 0.004$ (by SVM in Archidona) to $R^2 = 0.079$ (by MLP in Málaga), highlighting the non-autocorrelation between precipitation values from the previous and following days. Comparing the results between B and C, in terms of RMSE, on average, approach C (RMSE = 4.359 mm/day) obtained slightly better performance than approach B (RMSE = 4.323 mm/day). However, in area 2, the use of approach B (RMSE = 4.986 mm/day) was significantly better than approach C (RMSE = 5.588 mm/day).

Finally, in Figures 5. 5, and 5. 6, all the RMSE, MBE, and R^2 values, from both areas and all models, are represented in a scatter plot.

Table 5. 6. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the first area (inland locations), using data from the target station in two different approaches, with the use of the previous and following day and the use of the two previous and two following days. The best values from each station are in bold.

Stations (Area 1)	Model	One Day (B)			Two Days (C)		
		RMSE [mm/day]	MBE [mm/day]	R^2	RMSE [mm/day]	MBE [mm/day]	R^2
La Higuera de Arjona	MLP	**4.409**	**−0.021**	**0.023**	**4.079**	**0.061**	**0.051**
	SVM	4.601	−1.218	0.008	4.348	−1.225	0.027
	RF	4.524	−0.880	0.020	4.224	−0.932	0.033
Jaen	MLP	3.875	−1.071	0.016	**4.423**	**−0.016**	**0.022**
	SVM	3.857	−1.039	0.018	4.613	−1.189	0.007
	RF	**3.785**	**−0.771**	**0.019**	4.583	−1.103	0.011
Linares	MLP	4.797	−1.378	0.015	4.455	−1.260	0.010
	SVM	4.754	−1.308	**0.019**	4.423	−1.202	0.010
	RF	**4.719**	**−0.940**	0.012	**4.371**	**−0.911**	**0.014**
Mancha Real	MLP	**3.246**	**0.128**	**0.047**	**3.288**	**0.305**	**0.012**
	SVM	3.450	−0.946	0.005	3.390	−0.842	0.002
	RF	3.386	−0.820	0.021	3.383	−0.788	0.003
Marmolejo	MLP	5.530	−1.459	0.012	5.396	−1.374	0.014
	SVM	5.501	−1.410	0.022	5.360	−1.307	0.015
	RF	**5.474**	**−0.947**	**0.014**	**5.235**	**−0.761**	**0.028**
Sabiote	MLP	3.992	−1.159	0.026	4.186	−1.114	0.008
	SVM	3.937	−1.091	0.030	4.155	−1.041	0.006
	RF	**3.893**	**−0.797**	**0.016**	**4.119**	**−0.910**	**0.010**
Torreblascopedro	MLP	4.658	−1.287	0.022	4.283	−1.204	0.011
	SVM	4.626	−1.236	0.027	4.263	−1.167	0.010
	RF	**4.539**	**−0.900**	**0.021**	**4.202**	**−0.802**	**0.015**
Mean values		4.359	−0.978	0.034	4.322	−0.894	0.037

Table 5. 7. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the second area (coastal locations), using data from the target station in two different approaches, with the use of the previous and following day and the use of the two previous and two following days. The best values from each station are in bold.

Stations (Area 2)	Model	One Day (B)			Two Days (C)		
		RMSE [mm/day]	MBE [mm/day]	R^2	RMSE [mm/day]	MBE [mm/day]	R^2
Antequera	MLP	**5.035**	**−0.246**	**0.027**	4.521	−1.246	**0.048**
	SVM	5.243	−1.296	0.021	4.480	−1.197	0.045
	RF	5.229	−1.221	0.005	**4.467**	**−1.163**	0.017
Archidona	MLP	4.108	−1.095	0.008	**4.109**	**−0.059**	**0.041**
	SVM	4.089	−1.012	0.004	4.328	−1.180	0.027
	RF	**4.083**	**−0.480**	**0.023**	4.252	−0.695	0.029
Cártama	MLP	**5.479**	−1.149	0.009	**5.235**	**0.239**	**0.040**
	SVM	5.550	−1.144	**0.027**	5.431	−1.132	0.021
	RF	5.631	**−0.896**	0.021	5.374	−1.054	0.024
Churriana	MLP	6.551	−1.314	**0.051**	6.849	−1.406	**0.017**
	SVM	6.449	−1.263	0.045	6.817	−1.367	0.009
	RF	**6.448**	**−1.148**	0.022	**6.781**	**−1.263**	0.012
Málaga	MLP	**5.028**	**0.294**	**0.079**	6.850	−1.324	0.044
	SVM	5.279	−1.028	0.023	6.765	−1.273	**0.056**
	RF	5.104	−0.884	**0.079**	**6.693**	**−1.182**	0.041
Pizarra	MLP	**5.152**	**0.253**	0.031	**5.871**	**0.014**	0.050
	SVM	5.266	−1.071	**0.044**	6.064	−1.205	**0.081**
	RF	5.267	−0.785	0.025	6.058	−1.074	0.021
Vélez-Málaga	MLP	**5.295**	**0.191**	**0.076**	**5.360**	**0.149**	**0.061**
	SVM	5.535	−1.198	0.047	5.544	−1.144	0.040
	RF	5.465	−1.046	0.052	5.489	−1.054	0.056
Mean values		5.299	−0.835	0.019	5.587	−0.934	0.015

3.3. Comparison of the Two Areas

In order to compare the results in the two different areas, Figure 5. 7 shows the RMSE, MBE, and R^2 performance values for these two kinds of locations (inland and coastal), using the best approach (data from neighbor stations). In terms of RMSE mean values, MLP outperformed RF and SVM. Besides, the models applied to the coastal locations under-performed, on average, in all cases and obtained higher variability across sites, rather than inland ones. In terms of MBE mean values, RF and LI obtained values very close to 0, whereas SVM was overestimated in most stations. Finally, in terms of R^2, the results by ML models were quite similar in both inland and coastal locations. However, the results of LI were significantly worse than ML in coastal sites, whereas SVM performed worse on in-land sites than on coastal.

Additionally, Table 5. 8 displays the maximum improvement, in terms of RMSE, R^2, and MBE, comparing ML to LI (using the first approach). In inland sites, the RMSE improvement ranged from 0.031 mm/day in Torreblascopedro to 0.263 mm/day in Marmolejo, as well as from 0.004 (Torreblascopedro) to 0.048 (Marmolejo), in terms of R^2. On the other hand, the upgrades in coastal sites ranged from RMSE = 0.076 mm/day and $R^2 = 0.012$ (in Pizarra) to RMSE = 1.475 mm/day and $R^2 = 0.25$ (in Archidona). Thus, coastal lo-cations significantly differed between linear interpolation and ML models for gap-filling daily rainfall. In contrast, in inland areas, the improvement was not substantial.

Thus, using empirical approaches (such as LI) to gap-fill daily rainfall data is not recommended, especially in coastal sites; the results are worse than ML, due to the effect of sea distance.

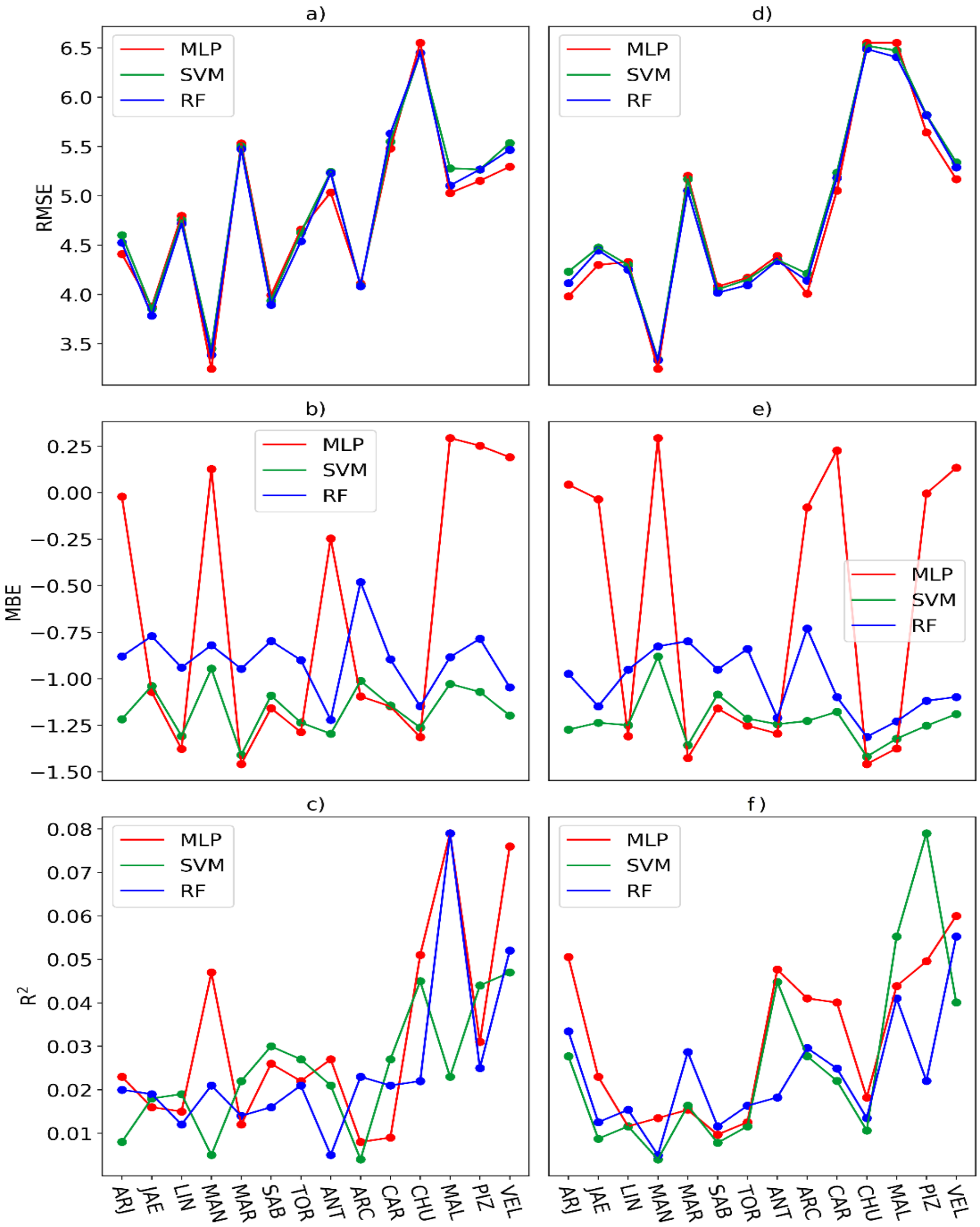


Figure 5. 6. RMSE, MBE, and R^2 values from testing dataset for all stations and models (MLP, SVM, and RF), using only precipitation data from the target station. A) RMSE using approach B, B) MBE using approach B, C) R^2 using approach B, D) RMSE using approach C, E)

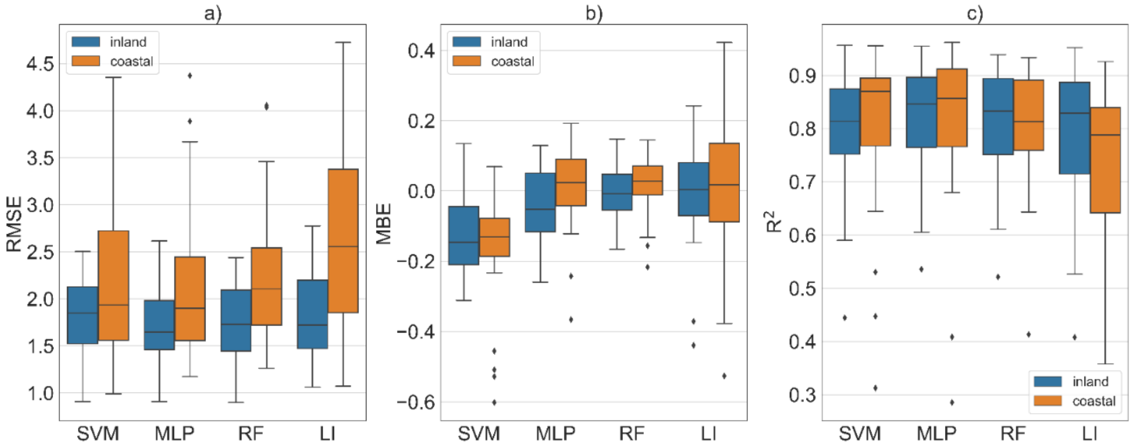


Figure 5. 7. RMSE (a), MBE (b), and R^2 (c) values of the different models (SVM, MLP, and RF) in the coastal and inland locations, using rainfall values from neighbor stations as inputs, where the minimum, the first interquartile (Q1), the median, the third interquartile (Q3), the maximum, and the outlier values are represented.

Table 5. 8. Best improvements between simple arithmetic averaging and the best ML model from each site for R^2 and RMSE. A positive value means that ML outperformed LI.

Station	RMSE (mm/day)	R^2
La Higuera de Arjona	0.139	0.025
Jaén	0.149	0.03
Linares	0.173	0.033
Mancha Real	0.19	0.041
Marmolejo	0.263	0.048
Sabiote	0.063	0.010
Torreblascopedro	0.031	0.004
Antequera	1.244	0.191
Archidona	1.475	0.25
Cártama	0.128	0.022
Churriana	0.781	0.086
Málaga	0.177	0.045
Pizarra	0.076	0.012
Vélez-Málaga	0.265	0.05

3.4. Seasonality Performance

In order to assess seasonal performance, the RMSE, MBE, and R^2 of all the stations and approaches, for the different evaluated models (SVM, MLP, and RF), are represented in Figure 5. 8. Regarding RMSE, summer, autumn, and spring obtained very similar average performances, whereas, in winter, the mean results were the worst. Moreover, summer obtained the narrowest interquartile range, but spring and winter got the more extensive range, with LI being the model with the worst range (the less confident between different stations) among all seasons and models. MLP, MLP, RF, and LI always performed with very similar average results, although LI had the widest interquartile range for all seasons. Besides, SVM always performed the worst, in terms of MBE. In terms of R^2, the highest mean values were carried out in winter, whereas the worst results were achieved in summer and spring. Regarding mean, all models performed with similar values during the same season.

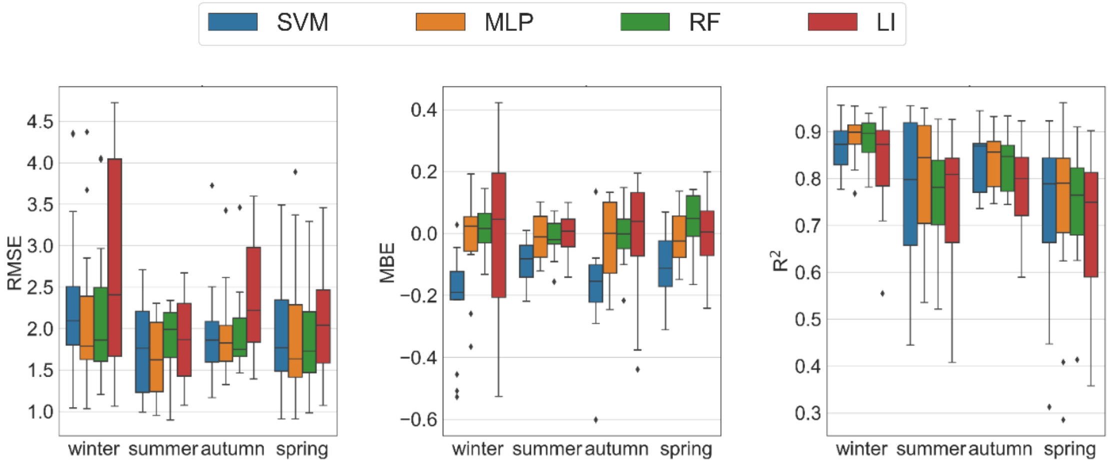


Figure 5. 8. Seasonality performance of the different models (SVM, MLP, and RF) in all the stations and approaches

Additionally, in Figures 5. 9, and 5. 10, the values predicted by the different ML models using the first approach are shown and compared to LI. In Figure 5. 9, the predictions from Torreblascopedro are plotted (the site that obtained the best performance, in terms of RMSE and R^2). In winter, all predictions are close to the 1:1 line, which denotes the excellent performance of this model during this season. The predictions were also close to the 1:1 line in spring and autumn, although the points were more dispersed than in winter. Finally, summer obtained the worst results, with the farthest points to the 1:1 line, especially with high rainfall values.

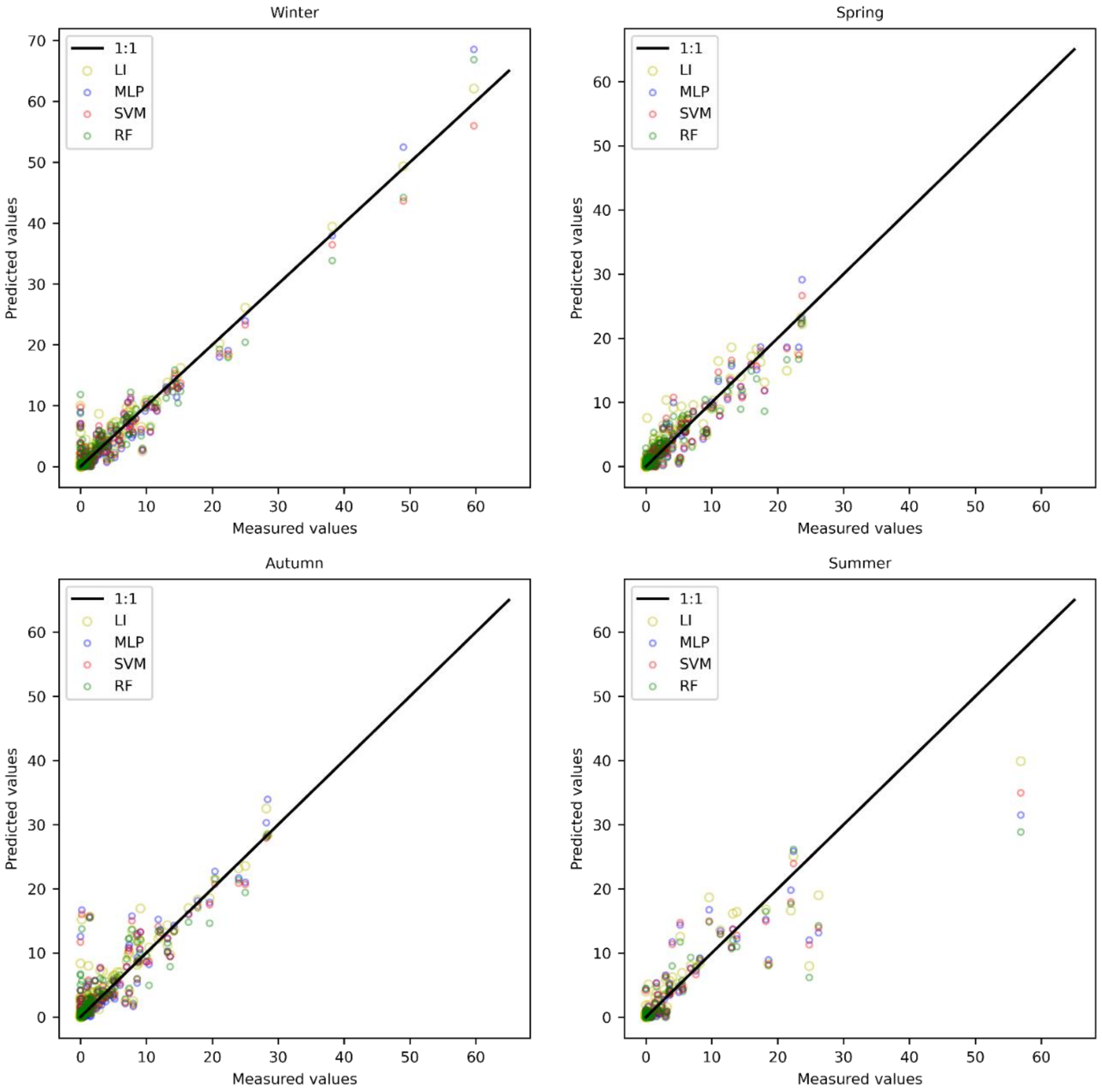


Figure 5. 9. . Scatter plot for predicted values in Torreblascopedro, using MLP, SVM, RF, and LI, during the different seasons

Finally, Figure 5. 10 plots the prediction rainfall values in Archidona. Spring obtained the best general predictions among all models, followed by autumn, summer, and winter, in this order. The highest differences between ML and LI were found in winter and autumn, where most LI predictions were farther from the 1:1 line.

Generally, summer obtained worse results than the rest of the seasons, due to the Mediterranean climate; during summers in Andalusia, precipitation is very occasional. They usually respond to local events, such as local torments. So, gap-filling rainfall data using neighbor stations with very different pluviometry makes models fail on those specific dates. Comparing the results between Torreblascopedro and Archidona, the most significant differences can be seen in winter, where LI performed much worse than ML approaches.

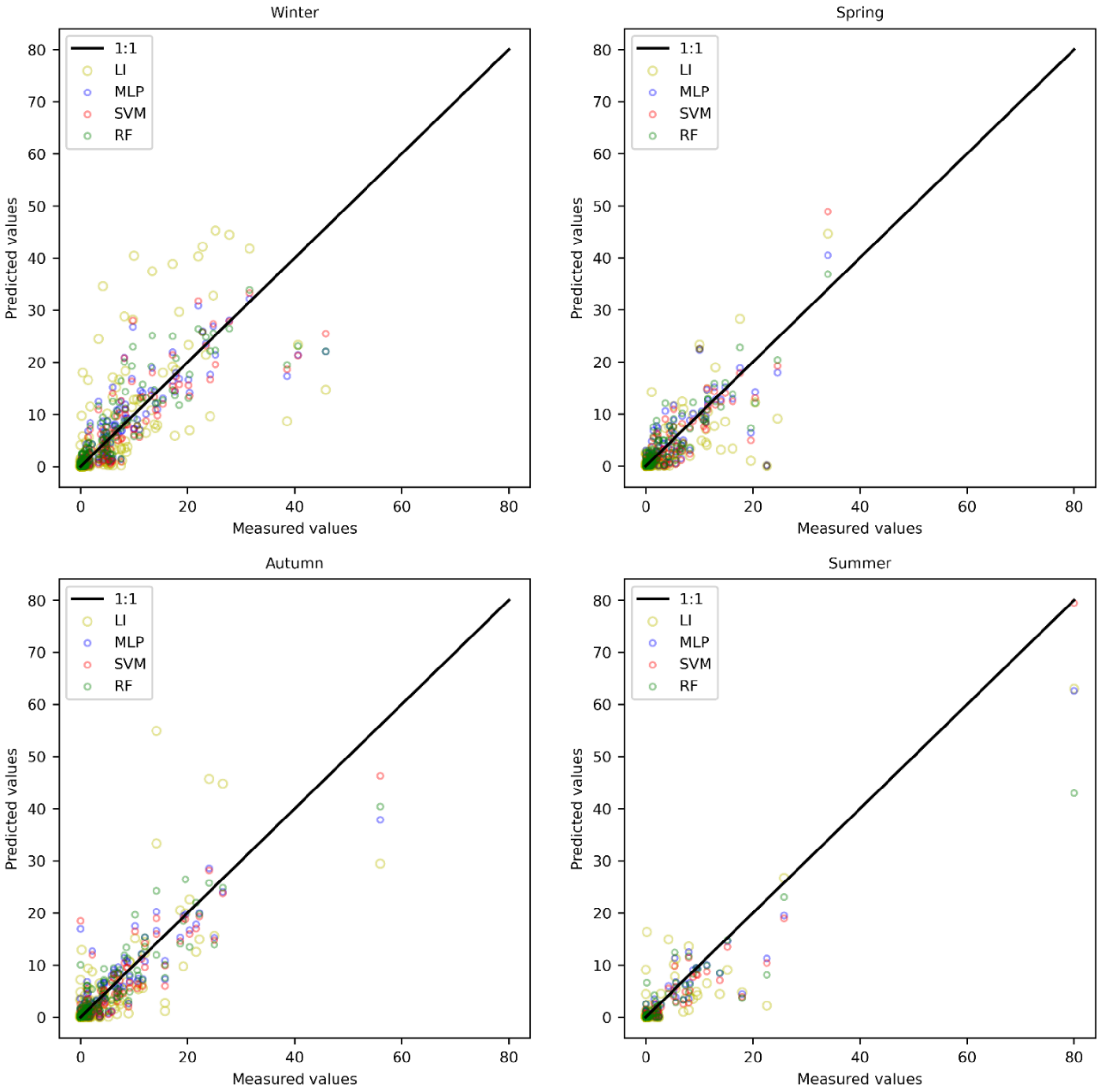


Figure 5. 10. Scatter plot for predicted values in Archidona, using MLP, SVM, RF, and LI, during the different seasons.

3.5. General Discussion

In terms of R^2, the results obtained in this work outperformed those obtained by Kim and Ryu [52] (Idaho, USA) using IDWM, OK, and GME, in conjunction with cluster analysis, having the best R^2 performance, with a value below 0.7 ($R^2 = 0.689$ or $R = 0.83$). Besides, the models developed in this work highly improved the RMSE and R^2 performance of Wuthiwongyothin [53] in Northern Thailand, using the K-means technique with the inverse distance weighting (IDW) and correlation coefficient weighting (CCW), where the mean R^2 values among all stations were below 0.6. Moreover, in terms of R^2, the values obtained by Sehad et al. [54] in North Algeria using multispectral MSG SEVIRI imagery were slightly worse, on average, than the obtained in this work, with a mean $R^2 = 0.7241$. However, in absolute terms, its developed model outperformed this work's best results ($R^2 = 0.921$ against $R^2 = 0.898$ in Torreblascopedro using SVM). Thus, ML models with neighbor station data located within a 50 km radius are highly recommended to gap-fill rainfall values in coastal locations, due to their accurate performance (among other approaches) in the different areas assessed along the Andalusia region, being the preferred use of neighbor stations, over the use of cluster analysis with stations located within a further radius distance. However, in inland sites, the performances carried out by ML against LI were not as significant as in coastal sites. Finally, in order to improve the state of the art of these approaches, future works could analyze the possibility of false alarms and missing rainfall cases using the models developed in this work.

4. Conclusions

Three different approaches were evaluated for gap-filling daily rainfall values: (A) the use of data from neighbor stations within 50km, (B) the use of one day before and ahead from the target station, and (C) the use of two days before and ahead from the target station. Fourteen different locations were evaluated from two areas, corresponding to inland and coastal sites. Additionally, three different ML models were assessed for this purpose: MLP, SVM, and RF. Daily large datasets of around 21 years were used (from 2000 to 2021), where 70% was used for training and a random 30% for testing purposes. Besides, 20% from the training dataset was used to find the fittest hyperparameters. Finally, a seasonality analysis was carried out. Based on the arisen results, no ML model significantly outperformed the rest, although MLP obtained the best results, or very close to them, in most locations. On the other hand, SVM had accurate performances in coastal sites, whereas the behavior was not so good at inland locations. RF behaved the opposite to SVM, having an accurate performance at inland locations and worse modeling at inland sites. Moreover, the first approach (the use of neighbor data) was notably better than the other approaches, with RMSE values below 2.0 mm/day and R^2 values above 0.85 in most stations. There were no significant seasonal differences in performance, in terms of RMSE and MBE values in winter, spring, and autumn, but the results obtained in summer were generally worse for all locations. Besides, coastal area location models performed slightly worse and with higher performance differences between ML and LI, in most sites and models, highlighting the differences in rainfall prediction efficiency, depending on the sea distance. In conclusion, it could be stated that the use of neighbor data with MLP is highly recommended as a rainfall gap-filling technique, rather than the use of data from the target station from the past and future. Moreover, when these works' results are compared to different papers' approaches using a cluster analysis from wider ranges, the use of closer stations (within a 50 km radius) obtained better results in terms of R^2.

Finally, due to the significant need to have a complete time series rainfall dataset on a daily basis and the increasing interest in installing low-cost wireless sensors (IoT), the models developed and assessed in this work can help with gap-filling datasets in this work near-real-time, thanks to the decreasing price of the low-cost automated weather stations using these new devices.

References

[1] Shen, R.; Huang, A.; Li, B.; Guo, J. Construction of a drought monitoring model using deep learning based on multi-source remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 2019, 79, 48–57, doi:10.1016/j.jag.2019.03.006.

[2] Fernández, A.J.; Molero, F.; Becerril-Valle, M.; Coz, E.; Salvador, P.; Artíñano, B.; Pujadas, M. Application of remote sensing techniques to study aerosol water vapour uptake in a real atmosphere. Atmos. Res. 2018, 202, 112–127, doi:10.1016/j.atmosres.2017.11.020.

[3] Astel, A.; Mazerski, J.; Polkowska, Z.; Namieśsnik, J. Application of PCA and time series analysis in studies of precipitation in Tricity (Poland). Adv. Environ. Res. 2004, 8, 337–349, doi:10.1016/S1093-0191(02)00107-7.

[4] Sayemuzzaman, M.; Jha, M.K. Seasonal and annual precipitation time series trend analysis in North Carolina, United States. Atmos. Res. 2014, 137, 183–194, doi:10.1016/j.atmosres.2013.10.012.

[5] Estévez, J.; Gavilán, P.; García-Marín, A.P.; Zardi, D. Detection of spurious precipitation signals from automatic weather stations in irrigated areas. Int. J. Climatol. 2015, 35, 1556–1568, doi:10.1002/joc.4076.

[6] Jiang, L.; Wu, J. Hybrid PSO and GA for Neural Network Evolutionary in Monthly Rainfall Forecasting; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7802.

[7] Cramer, S.; Kampouridis, M.; Freitas, A.A.; Alexandridis, A.K. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. Expert Syst. Appl. 2017, 85, 169–181, doi:10.1016/j.eswa.2017.05.029.

[8] Teegavarapu, R.S.V.; Chandramouli, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J. Hydrol. 2005, 312, 191–206, doi:10.1016/j.jhydrol.2005.02.015.

[9] Barrios, A.; Trincado, G.; Garreaud, R. Alternative approaches for estimating missing climate data: Application to monthly precipitation records in south-central Chile. For. Ecosyst. 2018, 5, 1–10, doi:10.1186/s40663-018-0147-x.

[10] McCuen, R.H. Hydrologic Analysis and Design, 3rd ed.; Pearson: New York, NY, USA, 2004; ISBN 978-0131424241.

[11] Bostan, P.A.; Heuvelink, G.B.M.; Akyurek, S.Z. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. Int. J. Appl. Earth Obs. Geoinf. 2012, 19, 115–126, doi:10.1016/j.jag.2012.04.010.

[12] Adhikary, S.K.; Muttil, N.; Yilmaz, A.G. Genetic Programming-Based Ordinary Kriging for Spatial Interpolation of Rainfall. J. Hydrol. Eng. 2016, 21, 04015062, doi:10.1061/(asce)he.1943-5584.0001300.

[13] Mair, A.; Fares, A. Comparison of Rainfall Interpolation Methods in a Mountainous Region of a Tropical Island. J. Hydrol. Eng. 2011, 16, 371–383, doi:10.1061/(asce)he.1943-5584.0000330.

[14] Simolo, C.; Brunetti, M.; Maugeri, M.; Nanni, T. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. Int. J. Climatol. 2010, 30, 1564–1576, doi:10.1002/joc.1992.

[15] Xia, Y.; Fabian, P.; Stohl, A.; Winterhalter, M. Forest climatology: Estimation of missing values for Bavaria, Germany. Agric. For. Meteorol. 1999, 96, 131–144, doi:10.1016/S0168-1923(99)00056-8.

[16] Teegavarapu, R.S.V.; Tufail, M.; Ormsbee, L. Optimal functional forms for estimation of missing precipitation data. J. Hydrol. 2009, 374, 106–115, doi:10.1016/j.jhydrol.2009.06.014.

[17] Teegavarapu, R.S.V. Estimation des données manquantes des précipitations en utilisant la proximité optimale d'imputation métrique base, la classification du plus proche voisin et méthodes d'interpolation à base de cluster. Hydrol. Sci. J. 2014, 59, 2009–2026, doi:10.1080/02626667.2013.862334.

[18] Huang, M.; Lin, R.; Huang, S.; Xing, T. A novel approach for precipitation forecast via improved K-nearest neighbor algo-rithm. Adv. Eng. Inform. 2017, 33, 89–95, doi:10.1016/j.aei.2017.05.003.

[19] Gorshenin, A.; Lebedeva, M.; Lukina, S.; Yakovleva, A. Application of Machine Learning Algorithms to Handle Missing Values in Precipitation Data. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin, Germany, 2019; Volume 11965, pp. 563–577.

[20] Bagirov, A.M.; Mahmood, A.; Barton, A. Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. Atmos. Res. 2017, 188, 20–29, doi:10.1016/j.atmosres.2017.01.003.

[21] Kajewska-Szkudlarek, J. Clustering approach to urban rainfall time series prediction with support vector regression model. Urban Water J. 2020, 17, 235–246, doi:10.1080/1573062X.2020.1760319.

[22] Estévez, J.; Bellido-Jiménez, J.A.; Liu, X.; García-Marín, A.P. Monthly Precipitation Forecasts Using Wavelet Neural Net-works Models in a Semiarid Environment. Water 2020, 12, 1909, doi:10.3390/w12071909.

[23] Partal, T.; Kişi, Ö. Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J. Hydrol. 2007, 342, 199–212, doi:10.1016/j.jhydrol.2007.05.026.

[24] Li, G.; Ma, X.; Yang, H. A hybrid model for monthly precipitation time series forecasting based on variational mode de-composition with extreme learning machine. Information 2018, 9, 177, doi:10.3390/info9070177.

[25] Filho, A.S.F.; Lima, G.A.R. Gap Filling of Precipitation Data by SSA—Singular Spectrum Analysis. In Proceedings of the Jour-nal of Physics: Conference Series; Institute of Physics Publishing: Bristol, UK, 2016; Volume 759, p. 012085.

[26] Sun, M.; Li, X.; Kim, G. Precipitation analysis and forecasting using singular spectrum analysis with artificial neural net-works. Clust. Comput. 2019, 22, 12633–12640, doi:10.1007/s10586-018-1713-2.

[27] Kim, S.; Hong, S.; Joh, M.; Song, S.K. DeepRain: ConvLSTM network for precipitation prediction using multichannel radar data. arXiv 2017, arXiv:1711.02316.

[28] Ha, J.-H.; Lee, Y.H.; Kim, Y.-H. Forecasting the Precipitation of the Next Day Using Deep Learning. J. Korean Inst. Intell. Syst. 2016, 26, 93–98, doi:10.5391/jkiis.2016.26.2.093.

[29] Chen, L.; Cao, Y.; Ma, L.; Zhang, J. A Deep Learning-Based Methodology for Precipitation Nowcasting with Radar. Earth Space Sci. 2020, 7, e2019EA000812, doi:10.1029/2019EA000812.

[30] Estévez, J.; Gavilán, P.; García-Marín, A.P. Spatial regression test for ensuring temperature data quality in southern Spain. Theor. Appl. Climatol. 2018, 131, 309–318, doi:10.1007/s00704-016-1982-8.

[31] Estévez Gualda, J.; Gavilán, P.; Giráldez, J.V. Guidelines on validation procedures for meteorological data from automatic weather stations. J. Hydrol. 2011, 402, 144–154, doi:10.1016/j.jhydrol.2011.02.031.

[32] Shanker, M.S.; Hu, M.Y.; Hung, M.S. Effect of data standardization on neural network training. Omega 1996, 24, 385–397, doi:10.1016/0305-0483(96)00010-2.

[33] Luna, A.M.; Lineros, M.L.; Gualda, J.E.; Giráldez Cervera, J.V.; Madueño Luna, J.M. Assessing the Best Gap-Filling Technique for River Stage Data Suitable for Low Capacity Processors and Real-Time Application Using IoT. Sensors 2020, 20, 6354, doi:10.3390/s20216354.

[34] Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. New machine learning approaches to improve reference evapotranspi-ration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agric. Water Manag. 2020, 245, 106558, doi:10.1016/j.agwat.2020.106558.

[35] Banadkooki, F.B.; Ehteram, M.; Ahmed, A.N.; Fai, C.M.; Afan, H.A.; Ridwam, W.M.; Sefelnasr, A.; El-Shafie, A. Precipitation forecasting using multilayer neural Network and support vector machine optimization based on flow regime algorithm taking into Account uncertainties of soft computing models. Sustainability 2019, 11, 6681, doi:10.3390/su11236681.

[36] Ortiz-García, E.G.; Salcedo-Sanz, S.; Casanova-Mateo, C. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. Atmos. Res. 2014, 139, 128–136, doi:10.1016/j.atmosres.2014.01.012.

[37] Nayak, M.A.; Ghosh, S. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. Theor. Appl. Climatol. 2013, 114, 583–603, doi:10.1007/s00704-013-0867-3.

[38] Aftab, S.; Ahmad, M.; Hameed, N.; Bashir, M.S.; Ali, I.; Nawaz, Z. Rainfall prediction in Lahore City using data mining tech-niques. Int. J. Adv. Comput. Sci. Appl. 2018, 9, 254–260, doi:10.14569/IJACSA.2018.090439.

[39] Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32, doi:10.1023/A:1010933404324.

[40]. Sukovich, E.M.; Ralph, F.M.; Barthold, F.E.; Reynolds, D.W.; Novak, D.R. Extreme quantitative precipitation forecast per-formance at the weather prediction center from 2001 to 2011. Weather Forecast. 2014, 29, 894–911, doi:10.1175/WAF-D-13-00061.1.

[41] Das, S.; Chakraborty, R.; Maitra, A. A random forest algorithm for nowcasting of intense precipitation events. Adv. Space Res. 2017, 60, 1271–1282, doi:10.1016/j.asr.2017.03.026.

[42] Wolfensberger, D.; Gabella, M.; Boscacci, M.; Germann, U.; Berne, A. RainForest: A random forest algorithm for quantitative precipitation estimation over Switzerland. Atmos. Meas. Tech. 2021, 14, 3169–3193, doi:10.5194/amt-14-3169-2021.

[43] Jiang, L.; Wu, J. Hybrid PSO and GA for neural network evolutionary in monthly rainfall forecasting. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2013; Volume 7802 LNAI, pp. 79–88.

[44] Wang, Y.; Liu, J.; Li, R.; Suo, X.; Lu, E. Precipitation forecast of the Wujiang River Basin based on artificial bee colony algo-rithm and backpropagation neural network. Alex. Eng. J. 2020, 59, 1473–1483, doi:10.1016/j.aej.2020.04.035.

[45] Kotthoff, L.; Thornton, C.; Hoos, H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic model selection and hy-perparameter optimization in WEKA. J. Mach. Learn. Res. 2017, 18, 1–5.

[46] Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1946–1956.

[47] Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.T.; Blum, M.; Hutter, F. Auto-sklearn: Efficient and robust automat-ed machine learning. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Mon-treal, QC, Canada, 7–12 December 2015; Volume 2, pp. 2962–2970.

[48] Hutter, F.; Kotthoff, L.; Vanschoren, J. Automated Machine Learning; The Springer Series on Challenges in Machine Learning; Hutter, F., Kotthoff, L., Vanschoren,

J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; ISBN 978-3-030-05317-8.

[49] Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. Assessing Neural Network Approaches for Solar Radiation Estimates Using Limited Climatic Data in the Mediterranean Sea. In Proceedings of the 3rd International Electronic Conference on Atmospheric Sciences (ECAS 2020), online, 16–30 November 2020.

[50] Borji, A.; Itti, L. Bayesian optimization explains human active search. Adv. Neural Inf. Process. Syst. 2013, 26, 55–63.

[51] Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proc. IEEE 2016, 104, 148–175, doi:10.1109/JPROC.2015.2494218.

[52] Kim, J.; Ryu, J.H. A heuristic gap filling method for daily precipitation series. Water Resour. Manag. 2016, 30, 2275–2294, doi:10.1007/s11269-016-1284-z.

[53] Wuthiwongyothin, S.; Kalkan, C.; Panyavaraporn, J. Evaluating Inverse Distance Weighting and Correlation Coefficient Weighting Infilling Methods on Daily Rainfall Time Series. SNRU J. Sci. Technol. 2021, 13, 71–79.

[54] Sehad, M.; Lazri, M.; Ameur, S. Novel SVM-based technique to improve rainfall estimation over the Mediterranean region (north of Algeria) using the multispectral MSG SEVIRI imagery. Adv. Space Res. 2017, 59, 1381–1394, doi:10.1016/j.asr.2016.11.042

Chapter 6

AgroML: an open-source repository to forecast reference evapotranspiration in different geo-climatic conditions using machine learning and Transformer-based models

Juan Antonio Bellido-Jiménez[1], Javier Estévez[1], Amanda P. García-Marín[1]

[1] Projects Engineering Area, Department of Rural Engineering, Civil Constructions and Projects Engineering, University of Córdoba, Spain

Abstract

Accurately forecasting reference evapotranspiration (ET_0) values is crucial to improve crop irrigation scheduling, allowing anticipated planning decisions and optimized water resources management and agricultural production. In this work, a recent state-of-the-art architecture has been adapted and deployed for multivariate input time series forecasting (Transformers) using past values of ET_0 and temperature-based parameters (28 input configurations) to forecast daily ET_0 up to a week (1 to 7 days). Additionally, it has been compared to standard machine learning models such as multilayer perceptron (MLP), random forest (RF), support vector machine (SVM), extreme learning machine (ELM), convolutional neural network (CNN), long short-term memory (LSTM) and two baselines (historical monthly mean value and a moving average of the previous seven days) in five locations with different geo-climatic characteristics in the Andalusian region, Southern Spain. In general, machine learning models significantly outperformed the baselines. Furthermore, the accuracy dramatically dropped when forecasting ET_0 for and horizon higher than three days. SVM, ELM, and RF using configurations I, III, IV, and IX outperformed, on average, the rest of the configurations in most cases. The best NSE values ranged from 0.934 in Córdoba to 0.869 in Tabernas, using SVM. The best RMSE, on average, ranged from 0.704 mm/day for Málaga and 0.883 mm/day for Conil using RF. In terms of MBE, most models and cases performed very accurately, with a total average performance of 0.011 mm/day. It was found a relationship in performance between the aridity index and the distance to the sea. The higher the aridity index at inland locations, the better results were obtained in forecasts. On the other hand, for coastal sites, the higher the aridity index, the higher the error. Due to the good performances and the availability as an open-source repository of these models, they can be used to accurately forecast ET_0 in different geo-climatic conditions, helping to increase efficiency in tasks of great agronomic importance, especially in areas with low rainfall or where water resources are limiting for the development of crops.

1. Introduction

The worldwide population is increasing to alarming values that will require almost 50% more food to meet the demand in 2050 [1]. Therefore, research into new methodologies to outperform agroclimatic forecasts (solar radiation, precipitation, or evapotranspiration) is a relevant task that allows the optimization of water resources management, the improvement of irrigation scheduling, and, indeed, contributes to the great challenge of increasing food production. Furthermore, it is significantly impactful in arid and semiarid areas such as the Andalusian region (Southern Spain), where crop water uses are elevated and the scarce precipitation is limiting growth and agricultural yield.

Crop evapotranspiration measures the crops' water demand, being affected by atmospheric parameters (such as temperature, wind speed, or solar radiation), specific crop types, soil characteristics, as well as management and environmental conditions. The evapotranspiration rate from a reference surface with no shortage of water is named reference evapotranspiration (ET_0), which studies the evaporative demand of the atmosphere independently of the surface, the crop type, its development stage, and the management practices. Its calculation can be accurately determined using physics-based methods like the FAO56-PM [2], which has been assessed globally in different climatic conditions and countries like Korea [3], Argentina [4], or Tunisia [5], among others. However, measuring all the required parameters (air temperature, relative humidity, wind speed, and solar radiation) is very costly in installation and maintenance. Moreover, Automated Weather Stations (AWS) usually contain non-reliable long-term datasets, mainly for wind speed and solar radiation, due to lack of maintenance or miscalibration [6]. These are the reasons why the geographical density of complete AWS is generally low, especially in rural areas and developing countries [7,8].

Therefore, developing new algorithms with fewer climatic input parameters is of high interest. In these terms, Hargreaves and Samani [9] introduced an empirical equation (HS) that uses maximum and minimum daily air temperature values (Tx and Tn, respectively) and extraterrestrial solar radiation (Ra). Different studies have assessed HS in different aridity conditions and countries, such as Iran [10], Italy [11], Bolivia [11], China [12], and others. Nevertheless, advances in computation during the last decades led to the application of new methodologies based on Artificial Intelligence (AI) with a very intensive computational cost. Thanks to the progress in CPU and GPU computation, the time spent training these models has dropped significantly, allowing scientists to apply them without needing a vast CPU/GPU farm and obtaining promising results in all sectors, especially agriculture. For example, Karimi et al. [13] evaluated the performance of Random Forest (RF) and other empirical methods to estimate ET_0 when several meteorological data were missing. RF surpassed the other models for temperature-based data availability when using Tx, Tn, Ra, and Relative Humidity (RH) as input features. Ferreira and da Cunha [14] assessed RF, Extreme Gradient Boosting (XGB), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN) to estimate daily ET_0 through different approaches using hourly temperature and relative humidity as features in different AWS in Brazil. CNN outperformed the rest of the models for most statistics and locations in both local and regional approaches. However, no optimization algorithm was used during hyperparameter tuning. Yan et al. [15] evaluated XGB to estimate daily ET_0 in two different regions (an arid and humid region) from China and seven meteorological input combinations using maximum and minimum daily temperature (Tx and Tn, respectively), extraterrestrial solar radiation (Ra), relative humidity (RH), wind speed (U_2) and sunshine hours (n). In order to tune the different hyperparameters, the Whale Optimization Algorithm (WOA) was used. Their results showed that using local and external (neighbor stations) datasets obtained even better performance than using only local data in some cases. Therefore, this strategy is very promising when there is a lack of long-term records. Wu et al. [16] studied the performance of Extreme Learning Machines (ELM) in different locations in China. They analyzed the use of the K-means clustering algorithm and the Firefly Algorithm (FFA) to estimate monthly mean daily ET_0 using Tx, Tn, Ra, and Tm (mean daily temperature). Nourani et al. [17] assessed Support Vector Regression (SVR), Adaptive Fuzzy Inference System (ANFIS), MLP, and Multiple Linear Regression (MLR) to forecast monthly ET_0 in Turkey, North Cyprus, and Iraq. Moreover, three ensemble methods were applied (simple averaging, weighted averaging, and neural ensemble) to outperform the performance and reliability of single modeling. The use of neural ensemble models highly outperformed single modeling in all cases, although simple and weighted averaging did not significantly perform better. Ferreira and da Cunha [18][18] evaluated the performance of daily ET_0 forecasts (up to 7 days) using CNN, Long Short-Term Memory (LSTM), CNN-LSTM, RF, and MLP using hourly data from different weather stations with heterogeneous aridity index characteristics in Brazil. In all cases, the use of the machine learning (ML) models outperformed the baselines, where CNN-LSTM performed the best in both local and regional scenarios using Tx, Tn, maximum and minimum relative humidity (RHx and RHn, respectively), wind speed, solar radiation (Rs), Ra, the day of the year (DOY) and ET_0 values from a lag window in the past (up to 30 days). In order to tune the different hyperparameters, a random search algorithm with 30 epochs was used.

In addition to these well-known and standard ML models, new architectures have been recently developed to deal with Natural Language Programming (NLP) problems with outstanding results, called Transformers [19]. The Transformer model is an Encoded-Decoder architecture based on a self-attention mechanism that looks at an input sequence and decides which timesteps are valuable. The promising results of Transformers have fostered their use on time series problems due to its apparent relationship. In both types of problems, words/parameter values are more or less meaningful based on their position. Therefore, several scientists have evaluated attention-based architectures in forecasting problems. For example, Wu et al. [20] proposed an Adversarial Sparse Transformer (AST) based on Generative Adversarial Networks (GAN). They assessed it to forecast five different public datasets: (I) an hourly time series electricity consumption dataset, (II) an hourly traffic level from the San Francisco dataset, (III) an hourly solar power production dataset, and (IV) an hourly time series dataset from M4 competition. Furthermore, [21] analyzed a Transformed-based architecture to forecast influenza-like illness (ILI), obtaining promising results. Finally, Li et al. [22] evaluated the performance of Transformers in time series forecasting using the same public datasets as Wu et al. [20] and obtained more accurate modeling with long-term dependencies.

This work is motivated by the need to minimize error in daily ET_0 forecasts, which is one of the main drawbacks in the reviewed literature, as well as the outstanding and promising performance of Transformers and Transformer-based models in different fields. Thereby, this work is the first one using a multivariate input Transformer-based architecture in order to forecast daily ET_0 (from one to seven days ahead). The development and assessment have been carried out using past values of ET_0 and temperature-based measured variables as features in five sites of Andalusia (Córdoba, Málaga, Conil, Tabernas, and Aroche) with different geoclimatic characteristics. Besides, standard ML models such as RF, MLP, SVR, ELM, CNN, and LSTM have been also evaluated in conjunction with Bayesian optimization to tune all their different hyperparameters. Thus, the main objectives of this work are a) to assess the performance of the proposed transformer model to forecast ET_0 and to compare it to standard ML models and two simple baselines (historical monthly mean value and mean of previous seven days); b) to evaluate different input feature configurations based on ET_0 past values and several temperature-based features to forecast ET_0 and c) to analyze the forecast efficiency depending on different geoclimatic characteristic sites.

2. Materials and Methods

2.1. Study area and dataset

Andalusia is located in the southwest of Europe, ranging from 37º to 39º N, from 1º to 7º W, and occupying an extension of 87 268 km². This work was carried out with data from five locations of Andalusia (Figure 6. 1), with different geoclimatic characteristics and representing a great variability in terms of UNEP aridity index [23] in this region (ranging from 0.555 – dry subhumid – in Aroche, to 0.177 – arid – in Tabernas). The coordinates and other characteristics of the AWS are reported in Table 6. 1. In contrast, in Table 6. 2, the minimum, mean, maximum, and standard deviation values of minimum, mean, and maximum daily air temperature (Tn, Tm, and Tx, respectively), relative humidity (RHn, RHm, RHx, respectively), wind speed (u_2), solar radiation (Rs) and reference evapotranspiration (ET_0) data are shown. The dataset used in this study belongs to the Agroclimatic Information Network of Andalusia (RIA), which can be downloaded at https://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController.

Figure 6. 1. Spatial distribution of Aroche, Conil, Córdoba, Málaga, and Tabernas in the Andalusia region, south of Spain

Table 6. 1. Geoclimatic characteristics of the locations assessed in this work (ARO – Aroche, CON – Conil de la Frontera, COR – Córdoba, MAG – Málaga and TAB – Tabernas). Time period from 2000 to 2018.

Site	Lon. (ºW)	Lat. (ºN)	Alt. (m)	Mean annual precipitation (mm)	UNEP Aridity index	Total available days
Aroche (ARO)	6.94	37.95	293	632	0.555 (dry-subhumid)	6399
Conil de la Frontera (CON)	6.13	36.33	22	470	0.479 (semiarid)	5868
Córdoba (COR)	4.80	37.85	94	589	0.462 (semiarid)	6397
Málaga (MAG)	4.53	36.75	55	434	0.366 (semiarid)	6438
Tabernas (TAB)	2.30	37.09	502	237	0.178 (arid)	6694

Table 6. 2. Minimum (Min), Mean, Maximum (Max), and Standard Deviation (Std) values of all the daily parameters measured: Maximum air temperature (Tx), Mean air temperature (Tm), Minimum air temperature (Tn), Maximum relative humidity (RHx), Mean relative humidity (RHm), Minimum relative humidity (RHn), Wind speed at 2m height (u_2), Solar radiation (Rs), Reference evapotranspiration (ET_0) at each location (Aroche -ARO-, Conil de la Frontera -CON-, Córdoba -COR-, Málaga -MAG- and Tabernas -TAB-) and for the whole dataset (2000-2018).

		Tx (ºC)	**Tm (ºC)**	**Tn (ºC)**	**RHx (%)**	**RHm (%)**	**RHn (%)**	**u_2 (m/s)**	**Rs (MJ/m^2day)**	**ET_0 (mm)**
ARO	Min	2.5	-0.2	-8.0	32.5	17.2	5.0	0.3	1.0	0.3
	Mean	23.2	16.1	8.9	89.5	65.9	39.0	1.2	17.8	3.2
	Max	44.0	34.1	24.9	100.0	100.0	100.0	5.8	34.3	8.7
	Std	8.1	6.8	5.6	11.2	17.7	19.4	0.5	8.8	2.0
CON	Min	6.4	0.7	-5.3	39.9	24.3	6.9	0.0	0.5	0.4
	Mean	23.0	17.4	12.1	89.3	72.5	50.5	1.3	18.0	3.2
	Max	41.3	31.9	26.9	100.0	99.6	97.1	7.9	31.7	9.3
	Std	5.7	5.2	5.3	9.0	12.3	14.6	1.0	7.8	1.8
COR	Min	3.3	0.0	-8.3	38.9	21.8	4.3	0.0	0.5	0.3
	Mean	24.6	17.4	11.0	86.8	64.1	37.3	1.6	17.7	3.6
	Max	45.7	34.7	27.6	100.0	100.0	100.0	7.5	33.2	9.6
	Std	8.5	7.3	6.2	12.0	18.1	19.3	0.7	8.5	2.3
MAG	Min	6.2	3.3	-4.2	36.0	19.4	4.6	0.0	0.3	0.4
	Mean	23.9	18.2	12.6	85.1	63.4	39.1	1.3	18.2	3.4
	Max	42.7	33.7	26.8	100.0	99.7	98.3	4.6	32.4	10.3
	Std	6.3	5.8	5.5	10.5	14.2	15.1	0.5	8.2	1.9
TAB	Min	4.3	-1.2	-8.2	28.6	16.8	2.8	0.1	0.2	0.4
	Mean	23.2	16.4	9.8	85.7	59.9	32.9	1.9	18.4	3.8
	Max	42.5	32.1	26.0	100.0	97.5	95.0	9.9	32.8	10.6
	Std	7.2	6.6	6.2	11.9	15.1	14.8	0.9	7.8	2.0

In this work, because the accurate estimation of ET_0 using limited meteorological data has been improved in recent years [14,24] and due to the high availability of temperature records, only temperature-based and ET_0 values from the past have been used as input features to forecast ET_0. Specifically, two different windows have been evaluated, the use of 15 and 30 days from the past. Moreover, several temperature-based variables have been calculated, such as $Energy_T$ (the area below the intraday temperature in a whole day), $Hourmin_{Tx}$ (the time when T_X occurs), $Hourmin_{Tn}$ (the time when Tn occurs), $Hourmin_{Sunset}$ (the time when sunset occurs), $Hourmin_{Sunrise}$ (the time when sunrise occurs), e_s (mean saturation vapor pressure), e_a (actual vapor pressure) and VPD (vapor pressure deficit), T_X-Tn, $Hourmin_{Sunset}$-$Hourmin_{Tx}$ and $Hourmin_{Sunrise}$-$Hourmin_{Tn}$. All the configurations assessed in this work contained Tx, Tn, DTR, and Ra as features due to their very high Pearson correlation (Figure 6. 2), and the rest of the configurations are selected based on their Pearson correlation values and the previous results on these same locations regarding ET_0 and solar radiation [24–26] estimations. The 27 different assessed configurations can be shown in Table 6. 3

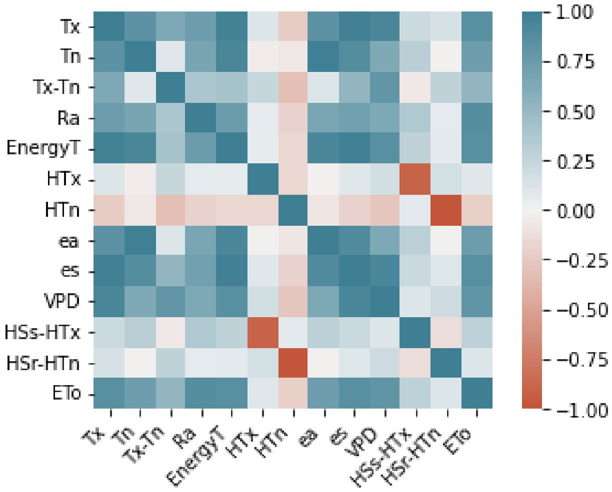


Figure 6. 2. Pearson correlation values of the assessed features in all the stations

Table 6. 3. Configuration table with all configurations. HT_X represents $Hourmin_{Tx}$, HTn represents $Hourmin_{Tn}$, HSs represents $Hourmin_{Sunset}$ and HSr represents $Hourmin_{Sunrise}$

Conf.	**T_X**	**Tn**	**DTR**	**R_a**	**$Energy_T$**	**e_a**	**e_s**	**VPD**	**HT_X**	**HTn**	**HSs – HT_X**	**HSr –HTn**	**ET_0**
I	X	X	X	X		X							X
II	X	X	X	X	X	X							X
III	X	X	X	X	X		X						X
IV	X	X	X	X	X								X
V	X	X	X	X	X						X		X
VI	X	X	X	X	X							X	X
VII	X	X	X	X	X					X			X
VIII	X	X	X	X	X	X			X				X
IX	X	X	X	X	X		X		X				X
X	X	X	X	X	X				X				X
XI	X	X	X	X	X				X		X		X
XII	X	X	X	X	X				X			X	X
XIII	X	X	X	X	X	X	X	X	X	X	X	X	X
XIV	X	X	X	X	X	X			X	X			X
XV	X	X	X	X	X		X		X	X			X
XVI	X	X	X	X	X				X	X			X
XVII	X	X	X	X	X				X	X	X		X
XVIII	X	X	X	X	X				X	X		X	X
XIX	X	X	X	X	X			X	X	X			X
XX	X	X	X	X	X			X	X				X
XXI	X	X	X	X	X			X					X
XXII	X	X	X	X			X						X
XXIII	X	X	X	X									X
XXIV	X	X	X	X							X		X
XXV	X	X	X	X								X	X
XXVI	X	X	X	X						X			X
XXVII	X	X	X	X					X				X

2.2. Preprocessing methodology

In machine learning applications, a vital prerequisite to guarantee accurate modeling is the use of reliable datasets. In this work, the control guidelines reported by Estévez et al. [6] have been followed to identify erroneous and questionable data from sensor measurements by applying different tests (range, internal consistency, step, and persistence) and spatial consistency test [27]. These quality assurance procedures have been successfully employed in different countries [4,28,29]. Afterward, the input and output matrixes had to be built depending on the number of lag days from the past (15 or 30), the features to use (up to 27 input configurations), and the number of days to forecast (up to 7 days). In Figure 6. 3, a mind map with all the possibilities is shown. It is worth noting that a MIMO (Multiple Input Multiple Output) approach was used in models that allowed it, whereas a Direct approach was considered in the others according to the results of Ferreira and da Cunha [18].

Consequently, using configuration 1 and 15 lag days as an example (see Figure 6. 4), the values from day to day-14 of Tx, Tn, Tx-Tn, Ra, e_a, and ET_0 are used as input features (a total of 90 values) for all the ML models (except for Transformers – see section 2.5.7), where Tx and Tn are directly given by AWS, Ra, and e_a can be calculated using Tx, Tn and the latitude, as it is stated by [2]. Finally, ET_0 is calculated using the well-known FAO56 PM method.

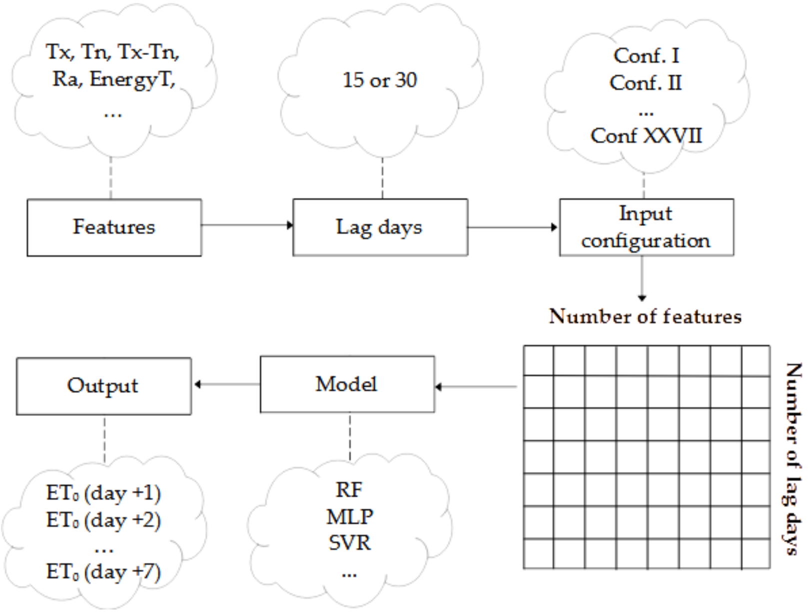


Figure 6. 3. Mind map of the matrix data structure

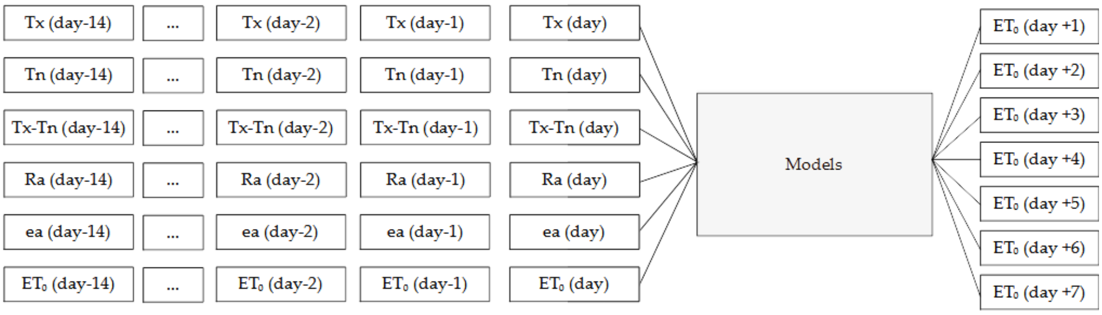


Figure 6. 4. Forecasting approaches using configuration 1 as an example

Later, in order to train, tune all the hyperparameters, and assess the final performance of the model, for each location, the dataset was split into training (70% of the entire dataset length), validation (20% of the training dataset length), and testing (30% of the entire dataset length) using a holdout technique. Next, the Bayesian optimization algorithm was used to tune all the hyperparameters (the hyperparameter space can be seen in Table 1 from Supplementary data). Eventually, after the best hyperparameter set was found, the final model was trained using the entire training dataset (70% of the entire dataset length) and evaluated using the testing dataset. Figure 6. 5 shows an overview of this methodology.

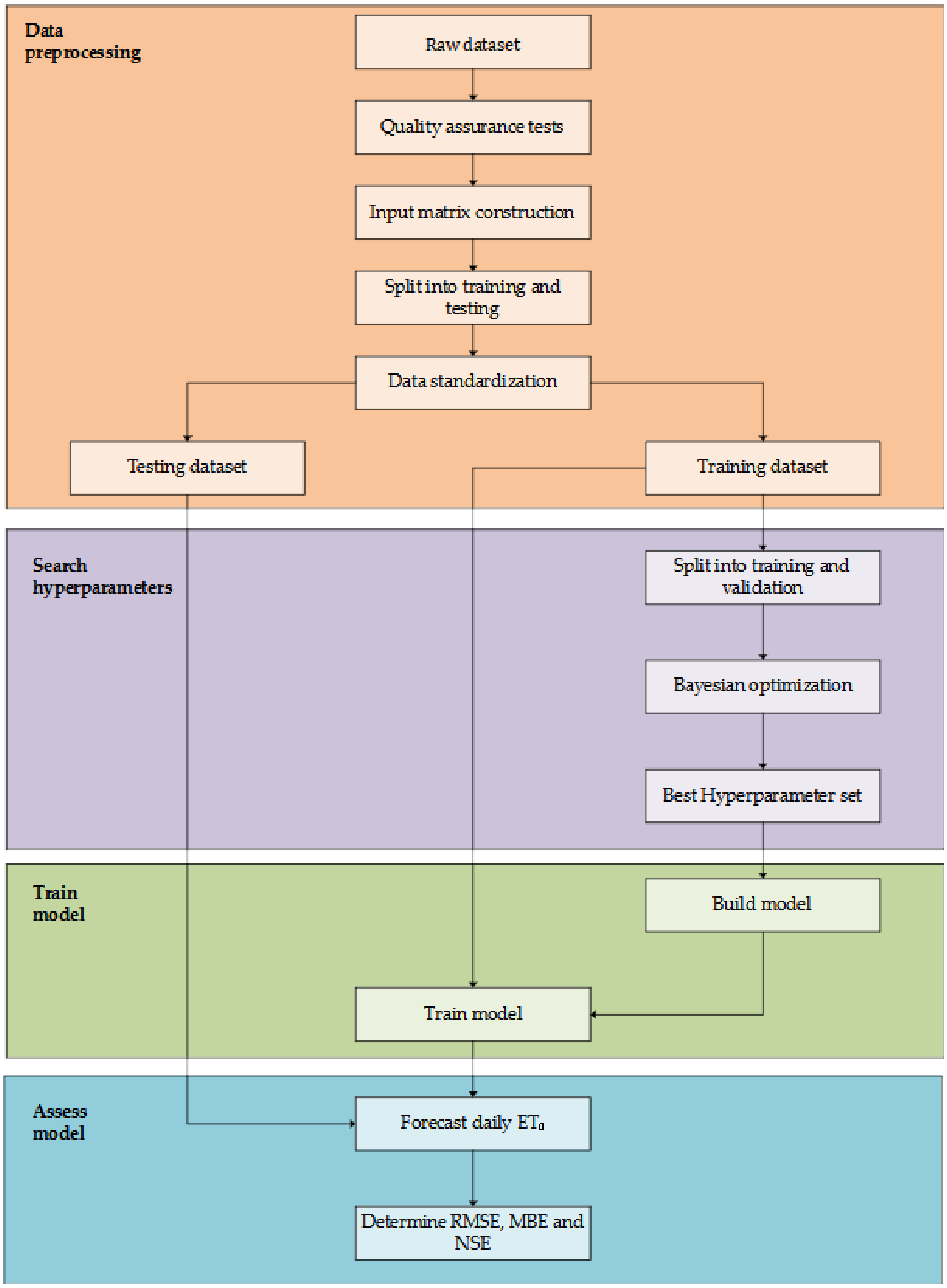


Figure 6. 5. Methodology flowchart

2.3. Reference evapotranspiration calculation

In this work, the ET_0 (FAO PM) values were used as input and target values. They were determined following the procedure of [2], and can be mathematically expressed as equation 6.1:

$$\mathrm{ET}_0 = \frac{0.408\Delta(\mathrm{Rn} - G) + \gamma \frac{900}{Tm + 273} u_2(\mathrm{es} - \mathrm{ea})}{\Delta + \gamma(1 + 0.34u_2)} \qquad \text{Eq 6.1}$$

where ET_0 is the reference evapotranspiration (mm day-1), 0.408 correspond to a coefficient (MJ^{-1} m^2 mm), Δ is the slope of the saturation vapor pressure versus temperature curve (kPa ºC^{-1}), Rn is the net radiation calculated at the crop surface (MJ m^{-2} day^{-1}), G is the soil heat flux density at the soil surface (MJ m^{-2} day^{-1}), γ is the psychrometric constant (kPa ºC^{-1}), Tm is the mean daily air temperature (ºC), u_2 is the mean daily wind speed at 2m height (m s^{-1}), es and ea are the saturation vapor pressure and the mean actual vapor pressure, respectively (kPa).

2.4. Baselines

In order to compare the performance of the developed models and configurations, it is crucial to have a baseline performance as a starting point. In this sense, two empirical baselines have been proposed in this work, following the methodology proposed by Ferreira and da Cunha [18]. In the first place, a moving average from the last 7 days is used. Secondly, the historical average monthly values from the training dataset were used to the corresponding forecast day.

2.5. Machine learning models

2.5.1. Multilayer Perceptron

The Multilayer Perceptron (MLP) is one of the most used agronomical and hydrological AI models [14,30,31]. Its popularity is based on its similarities to neurons in the biological nervous system, easy coding, and promising results in most cases. They are structured in three kinds of layers, the input and output layers, representing the inputs and outputs of the model, respectively, and the hidden layers, where all the neurons are located. The neurons work together to create stimuli (reference evapotranspiration forecast values) based on different inputs (the input matrix containing features from the past). A back-propagation algorithm makes the neurons learn (automatically update all weights and biases) and improve every mini-batch every epoch. A single neuron architecture can be seen in Figure 6. 6.

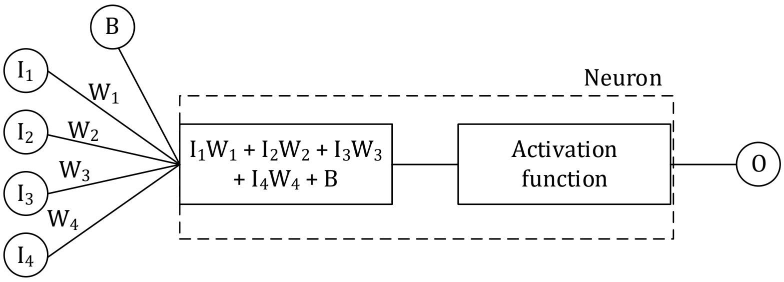


Figure 6. 6. Single neuron architecture. I1, I2, I3, and I4 represent the inputs of the neuron, W1, W2, W3, and W4 correspond to the weights of every input, B is the bias, and O represents the output of the neuron after passing through an activation function

2.5.2. Extreme Learning Machine

Extreme Learning Machine models (ELM) were first introduced by Huang et al. [32] as a single hidden layer feed-forward neural network with the following main characteristics: I) the input weights and biases are randomly generated and II) the output weights and biases are analytically determined. As a result, these models do not require any training process and have a meager computational cost, with promising results in ET_0 [24,33,34]. However, on the other hand, when the model is working with massive datasets, the amount of Random Access Memory (RAM) required is enormous.

2.5.3. Support Vector Machine for regression

Support Vector Machine (SVM) models for regression tasks, also known as Support Vector Regression (SVR) models, are supervised AI models based on a different functionality than neuron-based architectures such as MLP and ELM. They search for the best hyperplane (and its margins) that contains all data points. Thus, it could be easily related to linear regression with the flexibility of defining how much error can be considered acceptable. Moreover, one of their most important features is the use of kernels to allow the model to operate on a high-dimensional feature space. SVMs can be mathematically expressed as a minimization problem of equation 6.2 with the constraints in equation 6.3.

$$MIN\left(\frac{1}{2}\,\|w\|^2 + C\sum_{i=1}^{n}|\xi_i|\right) \geq 0 \qquad \text{Eq. 6.2}$$

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i| \qquad \text{Eq. 6.3}$$

where wi corresponds to the weight vector, x_i to the input vector, y_i to the output vector, ε represents the margins, ξ represents the deviation of values to the margins, C is a coefficient to penalize deviation to the margins and n the length of the training dataset. For further details, the work of [35] can be consulted.

2.5.4. Random Forest

A Random Forest (RF) is composed of the conjunction of multiple tree-based models in order to improve the overall result (ensemble model). The general idea is that different models are trained on different data samples (bootstrap) and feature sets. Instead of searching for the best features when splitting nodes, it searches among a random subset of the features. Thus, it results in greater diversity and better final performance.

2.5.5. Convolutional Neural Network

Convolutional Neural Network (CNN) models were first developed for image classification problems, where the convolution algorithm captures local patterns to learn a representation of figures to classify them. Moreover, this process can be extrapolated to 1D sequences of data such as time-series datasets. One of the advantages of using convolutions is that they can obtain local features relationships without the requirement of an extensive preprocessing method and obtaining outstanding results in ET_0 [14,36,37] and in other agro-climatic parameters [25,38,39].

Typically, such CNNs are composed of three layers, the convolutional layer, the pooling layer, and a fully connected layer. The convolutional layer is used to extract local relationships between the different features and timesteps. The pooling layer is added after the convolutional layer, and it gradually reduces the feature map. Finally, a fully connected layer is used to forecast the seven days horizon ET_0 values (in this work). For further details, the work of Aloysius et al. [40] can be reviewed.

2.5.6. Long Short-Term Memory

Long Short-Term Memory (LSTM) models were first introduced by Hochreiter et al. [41] as a Recurrent Neural Network (RNN) based model that could deal with long-term dependencies and address the vanishing gradient problem. In order to control the information flow, the LSTM block contains an input gate, an output gate, a forget gate, a cell state, and a hidden state. The gates are in charge of deciding which information is allowed on the cell state, so whether a piece of information is relevant to keep or forget during training. The cell and hidden state can be seen as the memory of the network, used to carry relevant information throughout the sequence.

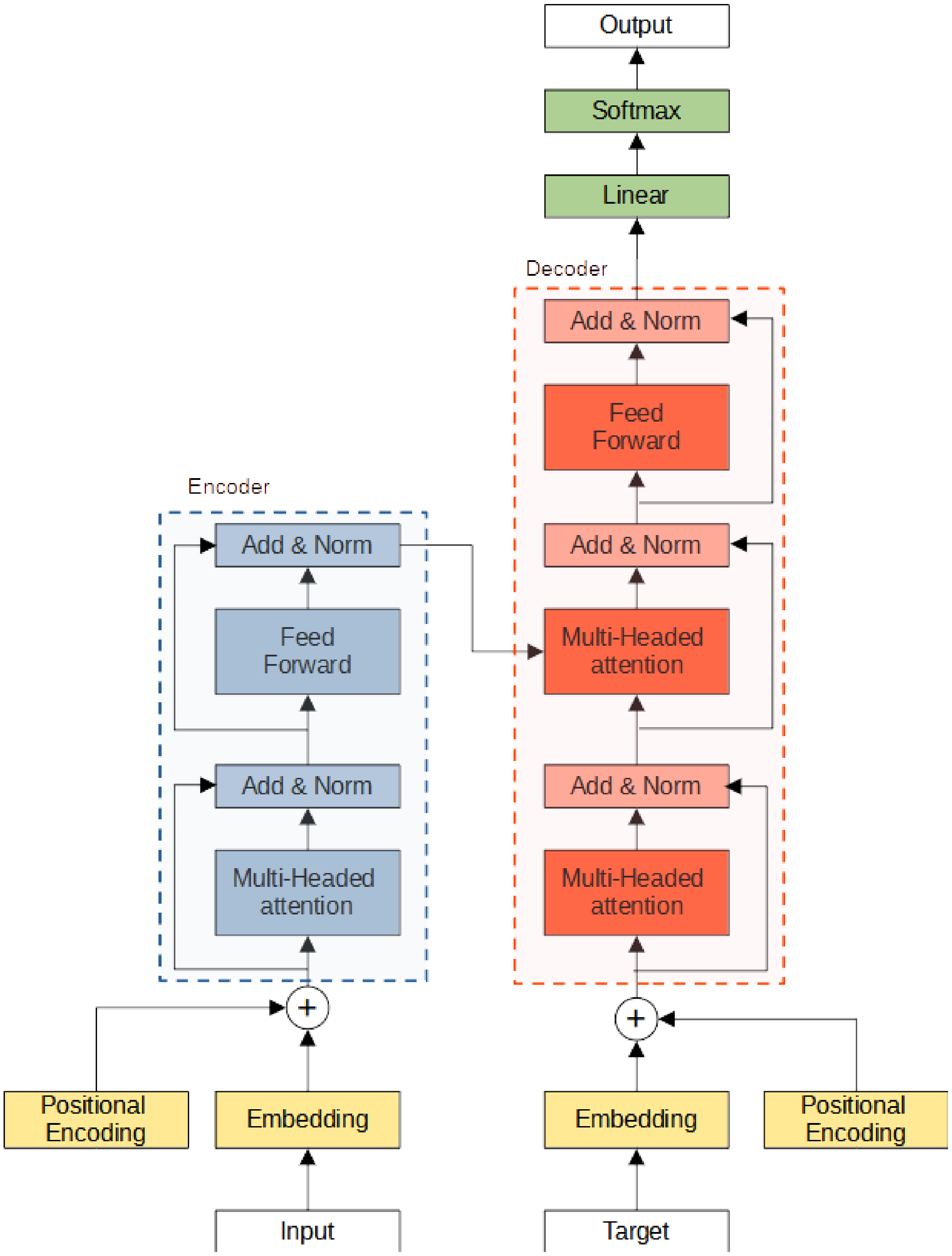


Figure 6. 7. Original Transformer architecture

2.5.7. Transformers

A new state-of-the-art architecture has been recently presented for NLP problems, the Transformers [19], see Figure 6. 7. One of the main motivations of Transformers is to deal with the vanishing gradient problem of LSTM when working with long sequences. Although LSTMs can theoretically propagate crucial information over infinitely long sequences, due to the vanishing gradient problem, they pay more attention to recent tokens and eventually forget earlier tokens. In contrast, Transformers use an attention mechanism, which learns what are the relevant subset of the sequences to accomplish the specific task. For a single head, the operation can be expressed as equation 6.4,

$$Attention(Q,K,V) = \ Softmax(\frac{Q\ K^T}{\sqrt{d_k}})V \qquad \text{Eq. 6.4}$$

where Q, K, and V represent the Query, Key and Value, respectively, as an analogy to a database, and dk corresponds to queries and keys' dimension. As stated by Yıldırım & Asgari-Chenaghlu(2021), the attention mechanism can be defined as : "*This can also be seen as a database where we use the query and keys in order to find out how much various items are related in terms of numeric evaluation. Multiplication of attention score and the V matrix produces the final result of this type of attention mechanism.".* Particularly, Transformers use a multi-head attention mechanism, that can be mathematically expressed as equation 6.5.

$$MultiHead(Q,K,V) = [Head_1, \ldots, Head_h]W_0 \qquad \text{Eq. 6.5}$$

where $Head_i$ is Attention(QW_i, KW_i, VW_i) and W are all the learnable parameter matrices.

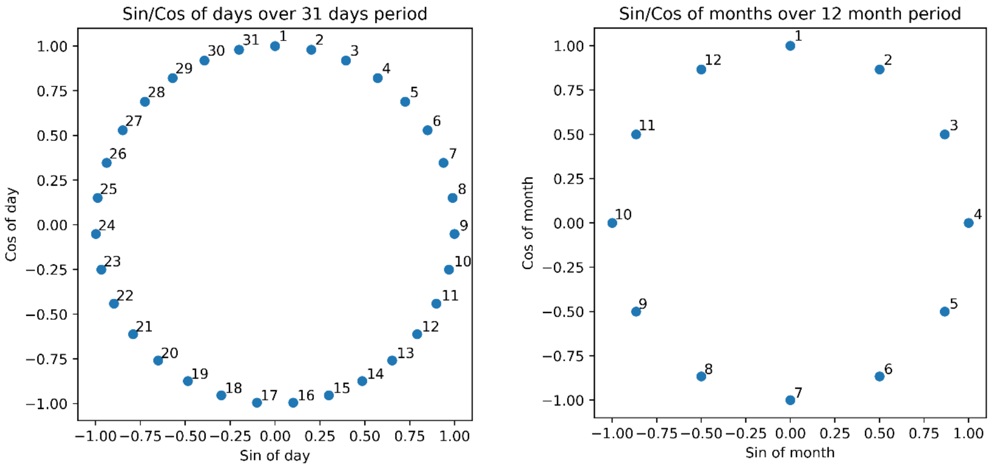
Generally, the Transformer is an encoder-decoder architecture. Considering translation task from English to Spanish, the encoder takes an input sequence ('*I am from Spain*') and maps it into a higher dimensional space using a Multi-Headed attention, an adding, a normalization, and a fully connected feed-forward layer. That abstract vector obtained in the encoder module is fed into the decoder, which uses it to obtain the translated sentence ('*Soy de España*'). It is worth noting that both encoder and decoder are composed of modules that can be stacked on top of each other multiple times. However, before carrying out any mathematical operation to the input data, it is required to convert words into numbers. The embedding layer is used for this purpose, transforming words into a vector of numbers that can be easily recognized by the model.

Another vital aspect to consider is the need for Transformers to learn temporal dependencies of the different timestamps through positional encoding because they do not inherently carry it out. In this work, the positional encoding was achieved using equations 6.6 and 6.7 for monthly and daily values (Figure 6. 8). In this way, 31st January and 2nd February are close, but 5th May and 26th July are not.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad \text{Eq. 6.6}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad \text{Eq. 6.7}$$

where pos represent the position, d_{model} is the input dimension, and i represents the index in the vector. It is worth noting that this temporal dependency information is shared with the rest of the models as new features in this work to make the comparison between models as fair as possible. Thus, new features are included to all configurations. For example, in configuration 1, the input features


Sin/Cos of days over 31 days period

Sin/Cos of months over 12 month period

a) b)

Figure 6. 8. Sine / Cosine positional encoding for 31 days in a month (a) and 12 months in a year (b)

The architecture used in this work can be seen in Figure 6. 9. It is based on the original Transformer architecture from Vaswani et al. [42] and the attention-based architecture of Song et al. [43]. Several aspects were modified. First, since the input data already has numerical values, the embedding layer was omitted. Then, the Positional encoding include new features to the input matrix instead of adding their values to the "embedded vector". Consequently, four more features were used in this model (sine and cosine positional encoding for days in a month, and sine and cosine positional encoding for months in a year). Finally, The SoftMax layer is also deleted because we are dealing with a regression problem (forecasting ET_0). Thus, the process of data in the proposed Transformer-based model can be described as follows. Firstly, the input matrix passes through a positional encoding mechanism. Then, the positional encoding features are added to the input matrix. Later, the data goes to an attention-based block containing multi-head attention, dropout, normalization, addition and feed-forward layers. Two different variations have been tested depending on the model used in the feed-forward layer: TransformerCNN where a convolutional approach has been used and TransformerLSTM where a LSTM approach has been implemented. Eventually, the processed data goes to a MLP model to carry out the regression task. The following works provide further details [44–47] and the code can be checked at AgroML github repository.

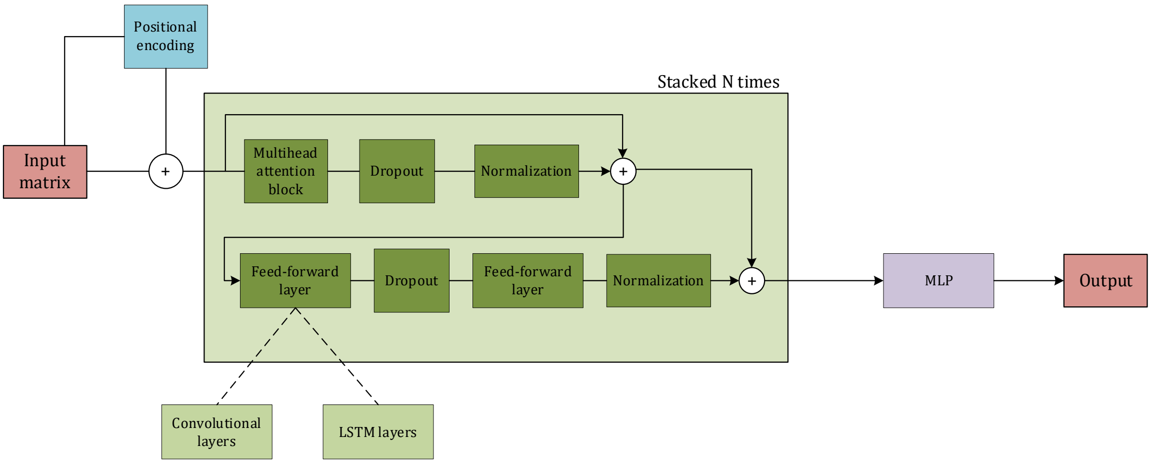


Figure 6. 9. The architecture of the proposed multi attention-based model

2.6. Bayesian optimization

The most critical aspect to obtain accurate performances in machine learning models is choosing the fittest hyperparameter set. The results could dramatically change from outstanding to very poor. A prevalent practice among the scientific community in agronomy and hydrology is using a trial-and-error approach [14,18,48], evaluating from dozens to hundreds of sets. However, it is not an efficient approach because the process is too slow if the hyperparameter space is large, spending a significant amount of time on non-promising configurations. Otherwise, if the hyperparameter space made to be small, you may obtain a suboptimal model. Several optimization algorithms have been assessed to solve this problem. For example, Particle-Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Genetic Algorithms (GA), Bayesian Optimization (BO), and Whale Optimization Algorithm (WOA), among others [31,49–51].

In this work, the BO algorithm has been proposed due to its high sample efficiency and popularity in Automated Machine Learning libraries such as Auto-Weka 2.0 [52], Auto-Keras [53], and Auto-Sklearn [54] and they can be consulted in Hutter et al. [55]. Part of its popularity is related to the close relationship to human behavior when carrying out this same process [56,57], where prior results are considered to choose the following set. BO is based on Bayes' theorem, and it can be explained using the following four steps: I) definition of the hyperparameter space; II) the algorithm first tries several random sets; III) the algorithm takes into account the previous assessed configuration sets when choosing the following one, a balancing between exploitation (it exploits regions that are known to have good performance) and exploration (choosing region with higher uncertainty) and evaluate it; IV) If the process has not finished yet, it goes to step 3.

In this work, BO has been implemented using Scikit-Optimize (gp_minimize) and Python 3.8. In all cases, this process was configured using 50 Bayesian epochs (80% of them are randomly chosen), selected after a trial and error algorithm among 50, 100, 150, and 200 bayesian epochs, the Mean Absolute Error (MAE) as the objective function and the rest of parameters as default. The hyperparameter space can be found in Table 1 from the Supplementary material and their results in Table 2.

Table 6. 4. Hyperparameter space for all the models assessed in this work. MLP – Multilayer Perceptron, RF – Random Forest, SVR – Support Vector Regression, ELM – Extreme Learning Machine, CNN –

Convolutional Neural Network, LSTM – Long Short-Term Memory, TransformerCNN – Transformer using CNN in the feed-forward layer, Transformer LSTM – Transformer using LSTM in the feed-forward layer,

Model	**Hyperparameter**	**Space**
MLP	Number of hidden layers	Up to 5
	Number of neurons	Up to 20
	Activation function	ReLU, Sigmoid and Tanh
	Maximum number of epochs	150
RF	Number of trees in the forest	From 10 to 1000
	Number of features to consider when looking for the best split	Sqrt, Log2
SVR	Kernel	Linear, Poly, Rbf, Sigmoid
	c	From 0.01 to 10
	Epsilon	From 0.01 to 10
ELM	Maximum number of neurons	150
	Activation function	Linear, Sigmoid, Tanh, Rbf_l1, Rbf_l2, Rbf_Linf
CNN	Number of convolutional block layers	From 1 to 5
	Number of convolutional layers in the block	From 1 to 5
	Number of filters	From 16 to 64
	Length of the convolutional window	From 1 to 20
	Length of the pooling window	From 1 to 5
	Type of pooling function	Average and Maximum
	Number of hidden layers	From 1 to 3
	Number of neurons	From 1 to 20
	Activation function	ReLU
	Optimizer function	Adam
	Maximum number of epochs	150
LSTM	Number of layers	From 1 to 5
	Number units	From 16 to 64
	Number of hidden layers	From 1 to 3
	Number of neurons	From 1 to 20
	Activation function	ReLU
	Optimizer function	Adam
	Maximum number of epochs	150
TransformerCNN	Number of attention heads	From 1 to 5

	Size of the head	From 1 to 5
	Number of filters	From 16 to 64
	Number of transformers blocks	From 1 to 3
	Length of the convolutional window	From 1 to 20
	Number of hidden layers	Up 1 to 3
	Number of neurons	Up to 20
	Maximum number of epochs	150
TransformerLSTM	Number of attention heads	From 1 to 5
	Size of the head	From 1 to 5
	Number of units	Fro 16 to 64
	Number of transformers blocks	From 1 to 3
	Number of hidden layers	Up 1 to 3
	Number of neurons	Up to 20
	Maximum number of epochs	150

Table 6. 5. Fittest hyperparameters for the best model and configuration at every location (All the hyperparameters name are the same as the variables in AgroML)

Model	Station	Conf	Hyperparameters
CNN	ALM04	7	n_conv_layers = 1 n_conv_base = 5 n_filters = 38 n_kernels = 20 type_pool = Max, pool_size = 2 n_hidden_layers = 1 n_neurons = 20 epochs = 150 activation = relu optimizer = adam
CNN	CAD05	26	n_conv_layers = 1 n_conv_base = 5 n_filters = 64 n_kernels = 20 type_pool = Avg pool_size = 4 n_hidden_layers = 3 n_neurons = 20 epochs = 149 activation = relu optimizer = adam
CNN	COR06	4	n_conv_layers = 1 n_conv_base = 3

			n_filters = 37 n_kernels = 18 type_pool = Avg pool_size = 2 n_hidden_layers = 2 n_neurons = 9 epochs = 145 activation = relu optimizer = adam
CNN	HUE06	4	n_conv_layers = 1 n_conv_base = 1 n_filters = 42 n_kernels = 20 type_pool = Avg pool_size = 2 n_hidden_layers = 1 n_neurons = 20 epochs = 150 activation = relu optimizer = adam
CNN	MAG01	22	n_conv_layers = 1 n_conv_base = 1 n_filters = 32 n_kernels = 10 type_pool = Avg pool_size = 5 n_hidden_layers = 1 n_neurons = 19 epochs = 149 activation = relu optimizer = adam
ELM	ALM04	1	activation = rbf_l2 neurons = 145
ELM	CAD05	2	activation = rbf_l2 neurons = 150
ELM	COR06	1	activation = rbf_l2 neurons = 150
ELM	HUE06	22	activation = rbf_l2 neurons = 150
ELM	MAG01	4	activation = rbf_l2 neurons = 150
LSTM	ALM04	6	n_layers = 1 units = 64

			n_hidden_layers = 1 n_neurons = 20 epochs = 150 activation = relu optimizer = adam
LSTM	CAD05	14	n_layers = 1 units = 59 n_hidden_layers = 1 n_neurons = 20 epochs = 150 activation = relu optimizer = adam
LSTM	COR06	16	n_layers = 1 units = 64 n_hidden_layers = 1 n_neurons = 20 epochs = 150 activation = relu optimizer = adam
LSTM	HUE06	8	n_layers = 1 units = 64 n_hidden_layers = 1 n_neurons = 15 epochs = 150 activation = relu optimizer = adam
LSTM	MAG01	22	n_layers = 1 units = 47 n_hidden_layers = 2 n_neurons = 15 epochs = 150 activation = relu optimizer = adam
MLP	ALM04	13	activation = relu optimizer = adam epochs = 59 neurons = [18, 16, 14, 9]
MLP	CAD05	14	activation = relu optimizer = adam epochs = 100 neurons = [20, 20]
MLP	COR06	7	activation = relu optimizer = adam

			epochs = 66 neurons = [13, 15, 18, 9, 14]
MLP	HUE06	2	activation = relu optimizer = adam epochs = 100 neurons = [20, 18, 13]
MLP	MAG01	25	activation = relu optimizer = adam epochs = 23 neurons = [20, 12]
RF	ALM04	3	n_estimators = 70 max_features = log2
RF	CAD05	16	n_estimators = 69 max_features = log2
RF	COR06	13	n_estimators = 100 max_features = sqrt
RF	HUE06	18	n_estimators = 75 max_features = log2
RF	MAG01	5	n_estimators = 60 max_features = log2
SVM	ALM04	21	kernel = rbf c = 5.680 epsilon = 0.103
SVM	CAD05	26	kernel = poly c = 2.203 epsilon = 0.077
SVM	COR06	13	kernel = linear c = 10.0 epsilon = 0.111
SVM	HUE06	11	kernel = rbf c = 0.805 epsilon = 0.052
SVM	MAG01	2	kernel = rbf c = 1.388 epsilon = 0.287
Transformer CNN	ALM04	5	head_size = 4 num_heads = 4 ff_dim = 29 num_transformer_blocks = 2 n_hidden_layers = 2 n_hidden_neurons = 17 n_kernel = 4 n_strides = 2

Transformer CNN	CAD05	20	head_size = 3 num_heads = 3 ff_dim = 29 num_transformer_blocks = 2 n_hidden_layers = 2 n_hidden_neurons = 16 n_kernel = 2 n_strides = 3
Transformer CNN	COR06	2	head_size = 3 num_heads = 5 ff_dim = 22 num_transformer_blocks = 2 n_hidden_layers = 1 n_hidden_neurons = 19 n_kernel = 4 n_strides = 5
Transformer CNN	HUE06	13	head_size = 5 num_heads = 3 ff_dim = 29 num_transformer_blocks = 3 n_hidden_layers = 1 n_hidden_neurons = 14 n_kernel = 2 n_strides = 1
Transformer CNN	MAG01	4	head_size = 5 num_heads = 4 ff_dim = 14 num_transformer_blocks = 3 n_hidden_layers = 1 n_hidden_neurons = 5 n_kernel = 5 n_strides = 2
Transformer LSTM	ALM04	3	head_size = 1 num_heads = 5 lstm_units = 64 num_transformer_blocks = 3 n_hidden_layers = 1 n_hidden_neurons = 20
Transformer LSTM	CAD05	18	head_size = 2 num_heads = 3 lstm_units = 56 num_transformer_blocks = 2 n_hidden_layers = 2

			n_hidden_neurons = 10
Transformer LSTM	COR06	8	head_size = 3 num_heads = 2 lstm_units = 52 num_transformer_blocks = 3 n_hidden_layers = 1 n_hidden_neurons = 12
Transformer LSTM	HUE06	10	head_size = 5 num_heads = 1 lstm_units = 64 num_transformer_blocks = 3 n_hidden_layers = 1 n_hidden_neurons = 20
Transformer LSTM	MAG01	4	head_size = 5 num_heads = 4 lstm_units = 49 num_transformer_blocks = 2 n_hidden_layers = 2 n_hidden_neurons = 16

2.7. Evaluation metrics

The model performances have been evaluated by using the following parameters: Mean Bias Error (MBE), Root Mean Square Error (RMSE), and the Nash-Sutcliffe model efficiency coefficient (NSE). The MBE, RMSE, and NSE are defined as equations 8, 9, and 10:

$$MBE = \frac{1}{m}\sum_{i=1}^{m} meas_i - pred_i \qquad \text{Eq. 6.8}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2} \qquad \text{Eq. 6.9}$$

$$NSE = 1 - \frac{\sum_{i=1}^{m}(meas_i - pred_i)^2}{\sum_{i=1}^{m}\left(meas_i - \overline{pred}\right)^2} \qquad \text{Eq. 6.10}$$

where meas and pred correspond to the observed and forecasted ET_0 values, respectively, n represents the number of records in the testing dataset and the bar denotes the mean.

3. Results and Discussion

It is worth noting that the code developed in this work is available on GitHub in the public repository called agroML, which can be found at https://github.com/Smarity/agroML. This new library focuses on helping scientists research state-of-the-art machine learning models, mainly focused on agronomy estimations and forecasts but easily extrapolated to other sectors and problems. It lets new scientists test these models on their datasets, and experienced scientists commit new features and architectures. The code has been programmed in standard Python using Tensorflow, Scikit-Learn, Scikit-Optimize, Pandas, and Numpy.

3.1. Baseline performance

Table 6. 6. RMSE values for ET0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean monthly value from the training dataset) and 6.7 show the RMSE and NSE performance for the baselines along the different forecast horizons (up to 1 week), where B1 refers to the moving average of last seven ET_0 values and B2 the use of mean historical monthly ET_0 values (mean ET_0 values for each month of the year). Generally, B2 outperformed B1 for all the forecast horizons except for one day ahead, where B1 performed better in all sites. Besides, B1 obtained the most accurate forecasts on the one-day ahead horizon, and it gradually dropped when the forecast horizon increased. In Aroche, the most humid site, the best performance in both RMSE and NSE values was obtained (NSE=0.9038 and RMSE=0.6390), followed by Córdoba, Málaga, Conil, and Tabernas (the aridest site), in this order. It suggests a relationship between the aridity index, distance to the sea, and the performance of models. In inland locations, the higher the aridity index, the fewer the forecasting errors. On the other hand, in coastal locations, it behaves right the opposite. The higher the aridity index and the far from the sea, the more precise the ET_0 modeling. Finally, Table 6.8 shows the MBE values for the different stations and forecast horizons. In this case, B1 outperformed B2 in most of the cases.

Table 6. 6. RMSE values for ET_0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean monthly value from the training dataset)

Location	**Baseline**	**Forecast Horizon**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
COR	B1	0.7551	0.8733	0.9365	0.9926	1.0172	1.0363	1.0644
	B2	0.8374	0.8374	0.8374	0.8374	0.8374	0.8374	0.8374
MAG	B1	0.7665	0.9084	0.9439	0.9632	0.9902	1.0140	1.0188
	B2	0.8143	0.8143	0.8143	0.8143	0.8143	0.8143	0.8143
TAB	B1	0.8515	0.9961	1.0451	1.0938	1.1075	1.1568	1.1628
	B2	0.9176	0.9176	0.9176	0.9176	0.9176	0.9176	0.9176
CON	B1	0.7987	1.0675	1.1950	1.2474	1.2404	1.2444	1.2778
	B2	0.9567	0.9567	0.9567	0.9567	0.9567	0.9567	0.9567
ARO	B1	0.6390	0.7882	0.8840	0.9337	0.9820	0.9901	1.0032
	B2	0.8027	0.8027	0.8027	0.8027	0.8027	0.8027	0.8027
Mean	B1	0.7622	0.9277	1.0009	1.0461	1.0675	1.0883	1.1054
	B2	0.8667	0.8667	0.8667	0.8667	0.8667	0.8667	0.8667

Table 6. 7. NSE values for ET_0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean daily monthly value from the training dataset)

Location	**Model**	**Forecast Horizon**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
COR	B1	0.8926	0.8564	0.8349	0.8145	0.8052	0.7978	0.7868
	B2	0.8680	0.8680	0.8680	0.8680	0.8680	0.8680	0.8680
MAG	B1	0.8376	0.7719	0.7538	0.7436	0.7290	0.7157	0.7129
	B2	0.8167	0.8167	0.8167	0.8167	0.8167	0.8167	0.8167
TAB	B1	0.8197	0.7531	0.7283	0.7023	0.6947	0.6671	0.6638
	B2	0.7906	0.7906	0.7906	0.7906	0.7906	0.7906	0.7906
CON	B1	0.8235	0.6844	0.6042	0.5684	0.5728	0.5695	0.5455
	B2	0.7465	0.7465	0.7465	0.7465	0.7465	0.7465	0.7465
ARO	B1	0.9038	0.8537	0.8160	0.7949	0.7732	0.7696	0.7636
	B2	0.8481	0.8481	0.8481	0.8481	0.8481	0.8481	0.8481
Mean	B1	0.8554	0.7849	0.7474	0.7247	0.7150	0.7039	0.6945
	B2	0.8140	0.8140	0.8140	0.8140	0.8140	0.8140	0.8140

Table 6. 8. MBE values for ET_0 forecast during seven forecast horizons and the two empirical baselines (B1 – using the average value from the last seven days – and B2 – using the mean daily monthly value from the training dataset)

Location	**Model**	**Forecast Horizon**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
COR	B1	-0.0002	-0.0001	-0.0001	0.0000	-0.0002	-0.0001	0.0007
	B2	0.1033	0.1033	0.1033	0.1033	0.1033	0.1033	0.1033
MAG	B1	0.0000	0.0002	0.0000	0.0000	-0.0008	-0.0016	-0.0015
	B2	0.0710	0.0710	0.0710	0.0710	0.0710	0.0710	0.0710
TAB	B1	0.0003	0.0003	0.0000	-0.0018	-0.0034	-0.0041	-0.0046
	B2	0.0972	0.0972	0.0972	0.0972	0.0972	0.0972	0.0972
CON	B1	0.0014	0.0047	0.0084	0.0117	0.0157	0.0198	0.0236
	B2	-0.0113	-0.0113	-0.0113	-0.0113	-0.0113	-0.0113	-0.0113
ARO	B1	0.0006	0.0011	0.0012	0.0021	0.0029	0.0036	0.0052
	B2	0.1787	0.1787	0.1787	0.1787	0.1787	0.1787	0.1787
Mean	B1	0.0004	0.0012	0.0019	0.0024	0.0028	0.0035	0.0047
	B2	0.0878	0.0878	0.0878	0.0878	0.0878	0.0878	0.0878

3.2. Analysis of ML performance

Table 6.9 shows the minimum, mean, and maximum NSE, RMSE, and MBE values for all the sites and models using two different lag intervals (15 and 30 days). Generally, in terms of NSE and RMSE, the use of 15 days slightly outperformed all the models using 30 lag days for almost all the cases. On the other hand, the MBE performance for all models, locations, and lag days was very similar to one another. Additionally, ML approaches highly outperformed the baselines, although the CNN and the transformer-based models gave the worst results in all sites. In Tabernas, the most arid site, in terms of NSE and RMSE, all the ML models surpassed the baselines performances. SVM obtained the best values (NSE=0.869 and RMSE=0.700 mm/day), followed very closely by RF (NSE=0.867 and RMSE=0.706 mm/day), which outperformed, on average, the rest of the models. On the other hand, the CNN model obtained the worst modeling for 30 lag days (NSE=0.423 and RMSE=1.438 mm/day). All the models obtained high mean MBE metrics, obtaining the highest MBE value (-0.974 mm/day) using CNN and 30 lag days. In Conil, the best values were obtained by SVM (RMSE=0.684 mm/day), RF (RMSE=0.703 mm/day) and ELM (RMSE=0.717 mm/day), in this order and for 15 lag days. In terms of NSE, these three models also gave the best performance on mean values and for 15 lag days, whereas the worst were obtained by CNN (NSE=0.520) for 30 lag days. In Córdoba, SVM and ELM using 15 lag days outperformed the rest of the models in both RMSE (0.605 and 0.614 mm/day) and NSE (0.934 and 0.932), respectively. Moreover, on average, the best results were obtained in Córdoba compared to the rest of the sites (NSE>0.85, RMSE<0.80 mm/day and MBE»0.0 mm/day). In Aroche, the most humid site, the NSE values ranged from 0.737 (CNN model) to 0.922 (SVM model) and the RMSE values ranged from 0.597 mm/day (SVM model) to 1.097 mm/day (CNN model). Finally, in Málaga, the results using 30 lag days were slightly better for all models. SVM and RF outperformed the rest of the models in terms of NSE (0.894 and 0.892, respectively) and RMSE (0.631 mm/day and 0.640 mm/day, respectively), whereas the worst results were obtained using CNN (NSE=0.409 and RMSE=1.499 mm/day) and LSTM (NSE=0.202 and RMSE=1.739 mm/day).

.

Table 6. 9. Minimum (Min.), mean and maximum (Max.) of NSE, RMSE, and MBE values for all locations (TAB – Tabernas, CON – Conil, COR – Córdoba, ARO – Aroche, MAG - Málaga) and models using two different lag days window (15 days and 30 days). T_CNN refers to Transformer using CNN in the feed-forward layer, while T_LSTM refers to Transformers using LSTM in this same layer

Station	Model	Lag days	NSE			RMSE			MBE		
			Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
TAB	CNN	15	0.71	0.778	0.862	0.723	0.916	1.050	0.001	0.123	0.484
		30	0.423	0.752	0.848	0.734	0.939	1.438	0	-0.026	-0.974
	ELM	15	0.794	0.82	0.86	0.727	0.825	0.885	0.043	0.082	0.126
		30	0.778	0.807	0.853	0.722	0.83	0.892	0	0.021	0.079
	LSTM	15	0.749	0.797	0.845	0.766	0.877	0.976	-0.003	0.088	0.236
		30	0.73	0.771	0.828	0.783	0.905	0.984	0	-0.009	-0.209
	MLP	15	0.769	0.81	0.854	0.743	0.848	0.936	0	0.046	0.265
		30	0.715	0.781	0.841	0.75	0.883	1.012	0	-0.029	-0.21
	RF	15	0.802	0.821	0.867	0.71	0.823	0.866	0.057	0.094	0.117
		30	0.799	0.819	0.859	0.706	0.805	0.85	0	-0.011	-0.033
	SVM	15	0.779	0.817	0.869	0.704	0.831	0.915	0	0.074	0.183
		30	0.746	0.812	0.862	0.7	0.818	0.955	0	-0.018	0.121
	T_CNN	15	0.742	0.789	0.84	0.779	0.893	0.989	0	0.1	0.324
		30	0.705	0.77	0.841	0.75	0.905	1.029	0	-0.017	-0.297
	T_LSTM	15	0.726	0.78	0.829	0.804	0.912	1.019	0.002	0.099	0.257
		30	0.699	0.765	0.831	0.775	0.916	1.040	0	-0.05	-0.312
CON	CNN	15	0.58	0.674	0.817	0.759	1.017	1.154	0	-0.037	-0.56
		30	0.303	0.52	0.724	0.889	1.164	1.409	0.002	-0.151	-0.706
	ELM	15	0.716	0.753	0.837	0.717	0.885	0.959	0	0	0.048
		30	0.635	0.697	0.779	0.796	0.927	1.021	-0.002	-0.057	-0.122

	LSTM	15	0.651	0.724	0.788	0.816	0.936	1.055	0	-0.029	-0.131
		30	0.378	0.552	0.706	0.919	1.126	1.326	0	-0.061	0.304
	MLP	15	0.579	0.709	0.808	0.778	0.959	1.160	0	-0.059	-0.26
		30	0.368	0.573	0.738	0.866	1.099	1.338	0.003	-0.153	-0.371
	RF	15	0.721	0.754	0.843	0.703	0.883	0.939	0.003	0.026	0.057
		30	0.667	0.704	0.799	0.759	0.915	0.967	-0.02	-0.054	-0.099
	SVM	15	0.64	0.752	0.851	0.684	0.885	1.065	0	-0.146	-0.25
		30	0.547	0.672	0.804	0.749	0.961	1.146	0.015	-0.235	-0.393
	T_CNN	15	0.561	0.679	0.8	0.794	1.008	1.184	0	-0.047	-0.225
		30	0.422	0.569	0.723	0.891	1.104	1.294	-0.001	-0.096	-0.451
	T_LSTM	15	0.57	0.674	0.746	0.895	1.018	1.177	0	-0.035	-0.166
		30	0.389	0.588	0.707	0.917	1.080	1.310	0	-0.082	-0.259
COR	CNN	15	0.818	0.882	0.929	0.63	0.808	1.011	0	0.056	-0.505
		30	0.522	0.853	0.913	0.67	0.873	1.592	0	0.035	1.003
	ELM	15	0.879	0.9	0.932	0.614	0.745	0.824	0	0.015	0.084
		30	0.848	0.874	0.909	0.686	0.813	0.896	-0.001	0.046	0.128
	LSTM	15	0.877	0.894	0.924	0.649	0.771	0.831	0	0.041	0.178
		30	0.835	0.865	0.902	0.713	0.841	0.932	0	0.027	0.193
	MLP	15	0.858	0.893	0.927	0.639	0.773	0.891	0	0.038	0.211
		30	0.801	0.858	0.908	0.69	0.86	1.029	-0.001	0.011	0.172
	RF	15	0.892	0.903	0.928	0.633	0.734	0.776	0.011	0.029	0.045
		30	0.87	0.883	0.912	0.674	0.783	0.826	0	0.015	0.033
	SVM	15	0.869	0.9	0.934	0.605	0.744	0.855	0	0.053	0.13
		30	0.832	0.875	0.914	0.667	0.809	0.942	0	0.064	0.167
	T_CNN	15	0.857	0.885	0.906	0.725	0.802	0.896	0.003	0.052	0.207

		30	0.815	0.855	0.892	0.749	0.87	0.988	0	0.023	-0.28
	T_LSTM	15	0.842	0.88	0.906	0.724	0.818	0.939	0	0.048	0.204
		30	0.824	0.859	0.885	0.773	0.859	0.965	0	0.037	0.23
ARO	CNN	15	0.799	0.851	0.913	0.624	0.816	0.951	0	0.106	0.436
		30	0.737	0.84	0.916	0.62	0.851	1.097	0.001	0.056	0.256
	ELM	15	0.85	0.874	0.917	0.609	0.751	0.823	-0.001	0.056	0.113
		30	0.853	0.878	0.918	0.613	0.744	0.819	0.02	0.082	0.141
	LSTM	15	0.823	0.86	0.912	0.627	0.792	0.892	0	0.068	0.196
		30	0.798	0.85	0.908	0.647	0.827	0.96	-0.002	0.038	0.22
	MLP	15	0.803	0.861	0.911	0.632	0.789	0.943	-0.001	0.079	0.288
		30	0.793	0.853	0.913	0.63	0.815	0.972	0	0.02	0.164
	RF	15	0.86	0.877	0.914	0.62	0.742	0.794	0.022	0.098	0.139
		30	0.855	0.883	0.92	0.606	0.73	0.814	0.009	0.047	0.07
	SVM	15	0.817	0.869	0.918	0.607	0.764	0.908	-0.003	0.136	0.2
		30	0.81	0.868	0.922	0.597	0.772	0.931	0.006	0.091	0.201
	T_CNN	15	0.802	0.845	0.902	0.664	0.834	0.945	0.002	0.099	0.281
		30	0.794	0.845	0.901	0.674	0.84	0.97	0	0.018	0.21
	T_LSTM	15	0.8	0.843	0.885	0.719	0.84	0.95	0	0.089	0.278
		30	0.78	0.838	0.882	0.736	0.859	1.001	0	0.042	0.238
MAG	CNN	15	0.734	0.8	0.871	0.681	0.847	0.98	0	0.046	0.311
		30	0.409	0.819	0.88	0.672	0.823	1.499	0	-0.003	1.113
	ELM	15	0.821	0.841	0.878	0.662	0.756	0.804	0	0.031	0.071
		30	0.841	0.857	0.884	0.663	0.736	0.777	-0.001	-0.04	-0.084
	LSTM	15	0.81	0.83	0.862	0.705	0.782	0.828	0	0.036	0.132
		30	0.202	0.84	0.872	0.695	0.773	1.739	0	-0.069	-1.052

MLP	15	0.773	0.823	0.872	0.678	0.798	0.904	0	0.036	0.195
	30	0.763	0.835	0.88	0.672	0.788	0.948	0	-0.048	-0.261
RF	15	0.832	0.849	0.882	0.651	0.738	0.778	0	0.027	0.044
	30	0.859	0.869	0.892	0.64	0.704	0.732	-0.02	-0.039	-0.061
SVM	15	0.797	0.843	0.885	0.643	0.75	0.855	0	0.049	-0.138
	30	0.814	0.858	0.894	0.631	0.731	0.839	0	-0.006	-0.094
T_CNN	15	0.741	0.809	0.853	0.727	0.829	0.967	0.001	0.009	0.198
	30	0.773	0.825	0.864	0.716	0.812	0.928	0.002	-0.097	-0.371
T_LSTM	15	0.768	0.801	0.835	0.771	0.846	0.916	0	0.001	-0.13
	30	0.787	0.827	0.852	0.749	0.808	0.897	0	-0.063	-0.247

In Figure 6. 10 and Figure 6. 11, the RMSE and NSE values for all forecasting predictions in the different sites were shown in a boxplot, respectively. Firstly, no significant performance distinctions were observed from the two approaches depending on the number of lag days (15 and 30 days). However, the first approach (15 lag days) slightly outperformed the second (30 lag days) on mean values, and more precision was observed (a lower interquartile range). Besides, the number of outliers having non-accurate modelings was much higher using the second approach. Then, as a general rule, using daily values from 15 days in the past is more recommended than using 30 days. Furthermore, regarding the efficiency of different models, SVM, RF and ELM were predominantly better than the rest of the models according to NSE and RMSE values, giving more precise results. In contrast, CNN and both Transformer models were at the bottom in the ranking. Finally, in Figure 6. 12 , the MBE results are plotted in a boxplot. The results were very accurate in both approaches and for all the models and sites, but CNN gave more outliers, especially using the 30 lag days approach.

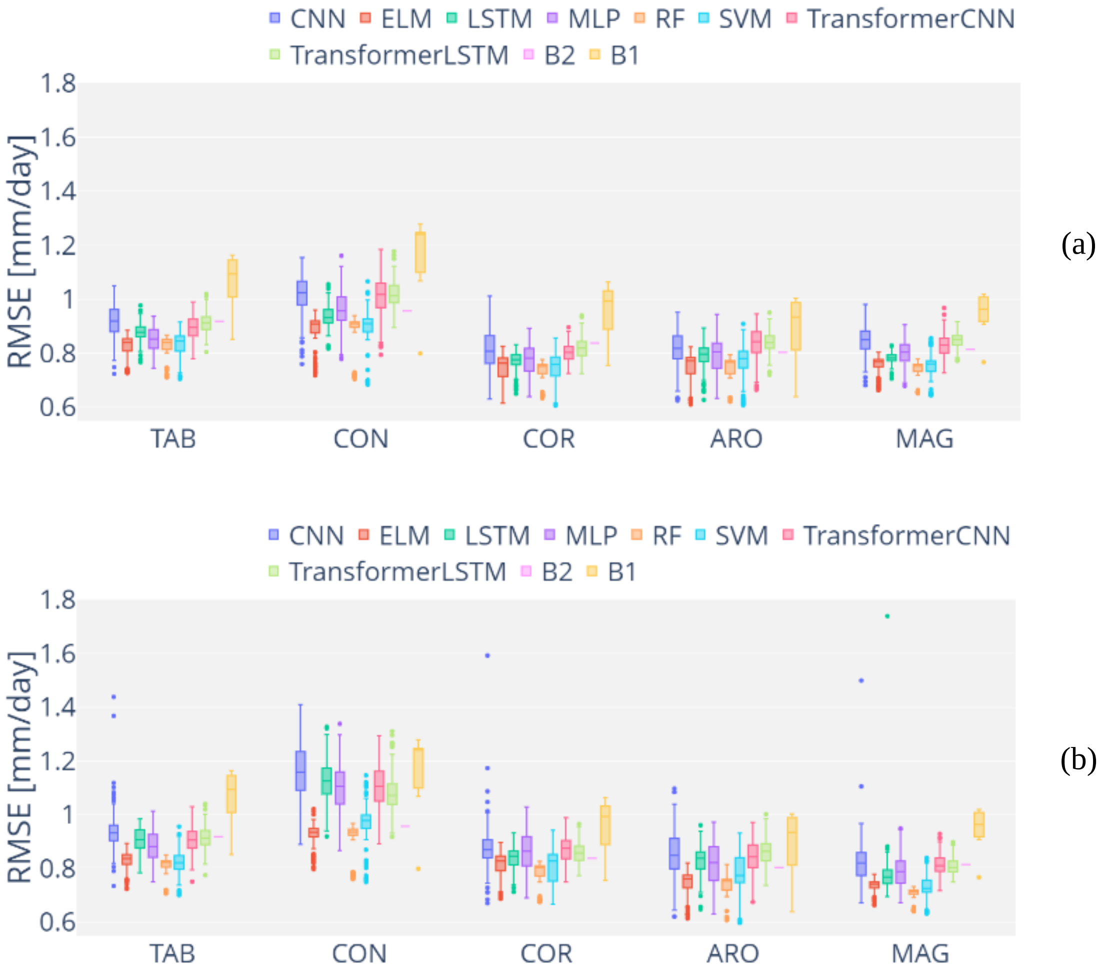


Figure 6. 10. Boxplot with RMSE values from all models and configurations in the different AWS, using 15 lag days (a) and 30 lag days (b)

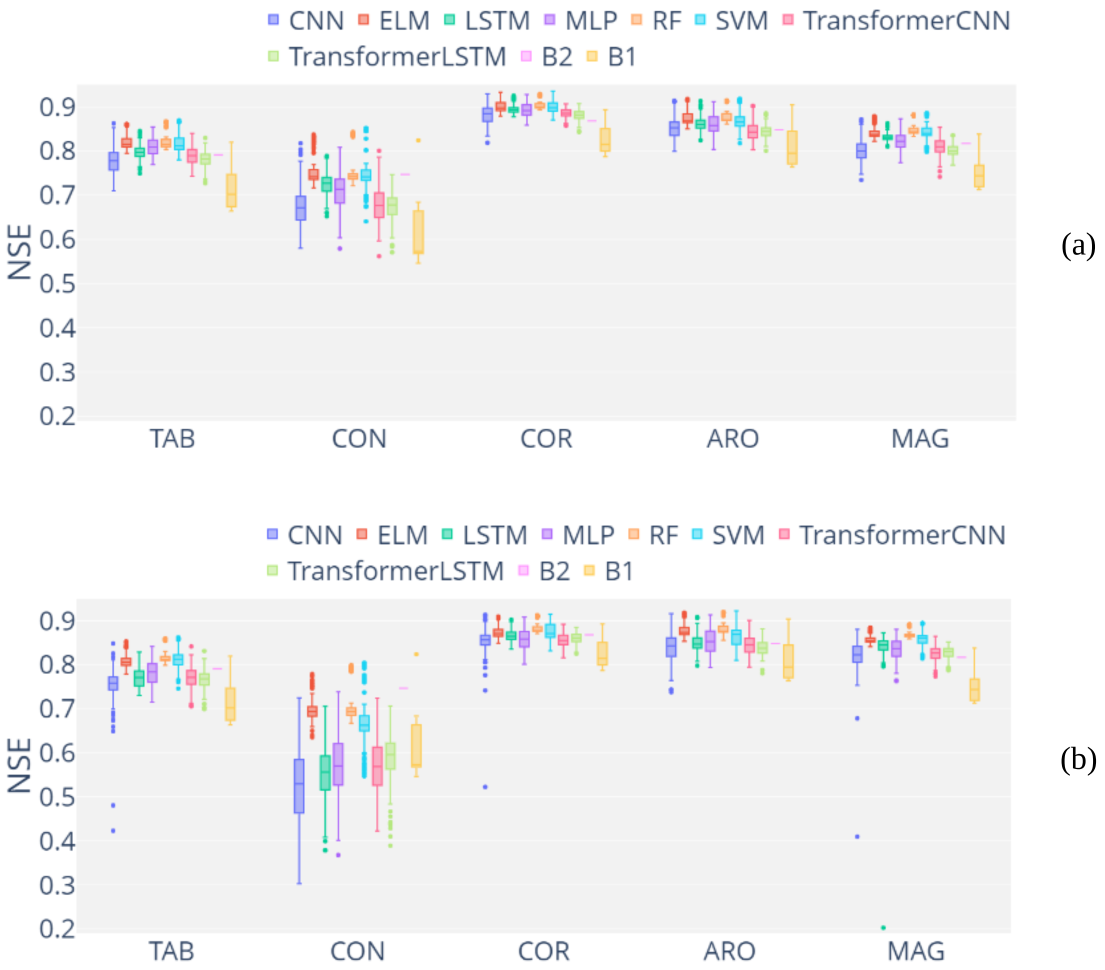


Figure 6. 11. Boxplot with NSE values from all models and configurations in the different AWS, using 15 lag days (a) and 30 lag days (b)

1

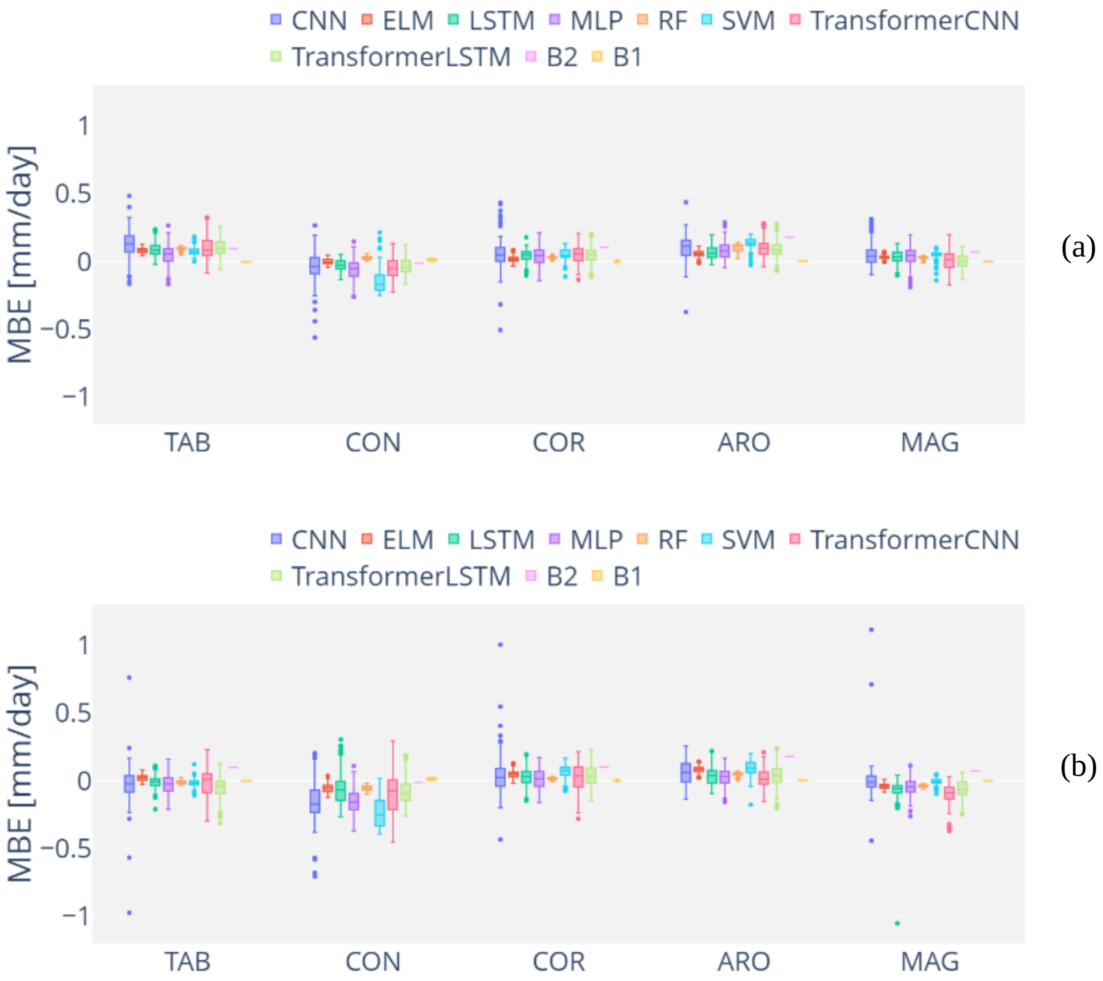


Figure 6. 12. Boxplot with MBE values from all models and configurations in the different AWS, using 15 lag days (a) and 30 lag days (b)

To further analyze these results, Figures 6.13, 6.14 and 6.15 show the best statistic values (NSE, RMSE and MBE, respectively) of all the models and sites for the different forecast horizons used. In terms of NSE (Figure 6.13), all ML models highly outperformed B1 and B2 in all the forecast horizons and locations, except for Conil. In Conil, only SVM, RF, and ELM outperformed both B1 and B2 in all cases. On the other hand, the Transformers, CNN and MLP models underperformed B1 and B2 for a horizon higher than 3 days. Regarding RMSE, the results were similar to Figure 6.14. However, a more significant improvement of ML models is appreciated for most models and horizons. In terms of MBE (Figure 6.15), B2 obtained significantly worse results in Aroche, Córdoba, Málaga and Tabernas, where ML performed very accurately in all cases. In Conil, there are no major differences in performance between all the models. Thereby, due to these results, it could be stated that the use of ML models to forecast ET_0 up to a week is highly recommended, especially SVM, RF, and ELM models. Generally, B1 highly outperformed B2 to forecast ET_0 values one day ahead, but its performance profoundly decreased for higher horizons, obtaining even worse results than B2. It denotes a low autocorrelation of daily ET_0 values but a higher relation with historical monthly values. Besides, SVM is generally on top of performance in terms of NSE and

RMSE, whereas regarding MBE, all models performed very accurately. Finally, it is worth noting that in Conil (a coastal site with an aridity index close to being a dry sub-humid climate), the best ML models (SVM, RF, and ELM) could not highly outperform B2 as in the rest of the locations when forecasting more than two days ahead, due to the effect of a close distance to the sea and the higher aridity index.

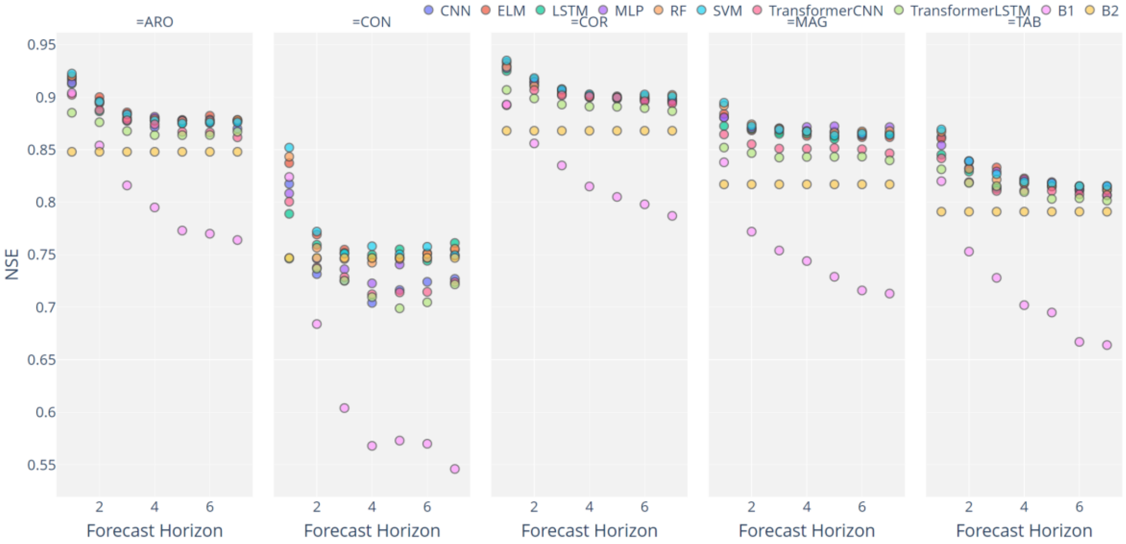


Figure 6. 13. Scatter plot with the best NSE value for each model and location.

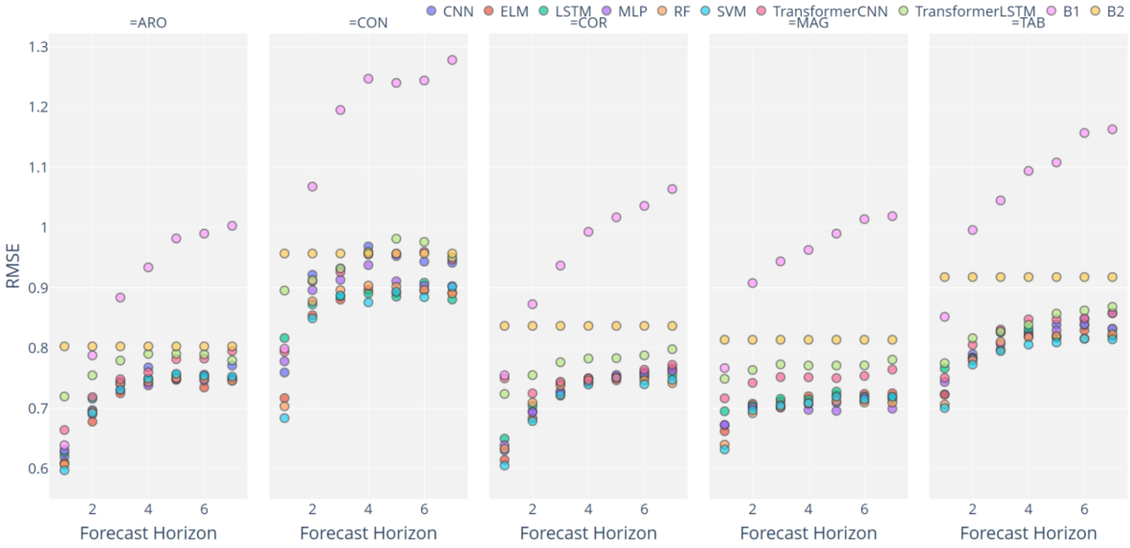


Figure 6. 14. Scatter plot with the best RMSE value for each model and location.

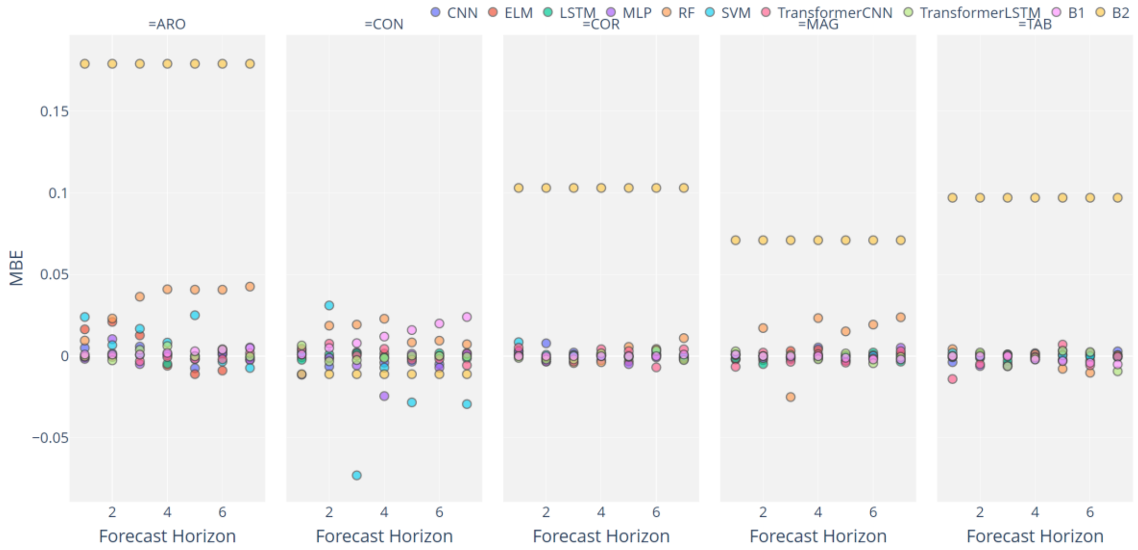


Figure 6. 15. Scatterplot with the best MBE value for each model and location.

3.3. Assessing the different configurations

In order to evaluate the performance of the different configurations at all locations, Table 6.10 shows the average and best RMSE values of each configuration in the different sites. In Tabernas, configurations III, XXII, IV, and IX obtained the most accurate results on mean, whereas configurations XVI, XII, and XXIV were the worst. In Conil, the best configurations in terms of mean RMSE were XXV, VI, and XX. Furthermore, configuration XXVI obtained the best value in absolute terms. On the other hand, configurations XIII, XI, and XII performed as the worst on average. In Córdoba, regarding mean values, configurations XVII, XXIV, and V were at the bottom, whereas configurations III, XXVII, and II were at the top of the ranking. In Aroche, configuration V obtained the lowest RMSE value (RMSE=0.598 mm/day). Besides, considering the mean values, all configurations obtained very similar performance, ranging from RMSE=0.764 mm/day (configuration I), followed closely by configurations IV (RMSE=0.764 mm/day), III (RMSE=0.767 mm/day), IX (RMSE=0.767 mm/day) and XXII (RMSE=0.768 mm/day), to RMSE=0.788 mm/day (configurations XIII and XVII). Thus, it could be stated that on mean, despite there being no significant differences in performances between the best and worse configuration, the use of configurations I, III, IV, and IX were more recommended.

Table 6. 10. Mean and minimum RMSE values (mm/day) for the different configurations at each location. The format is: mean (minimum). The best values are in bold

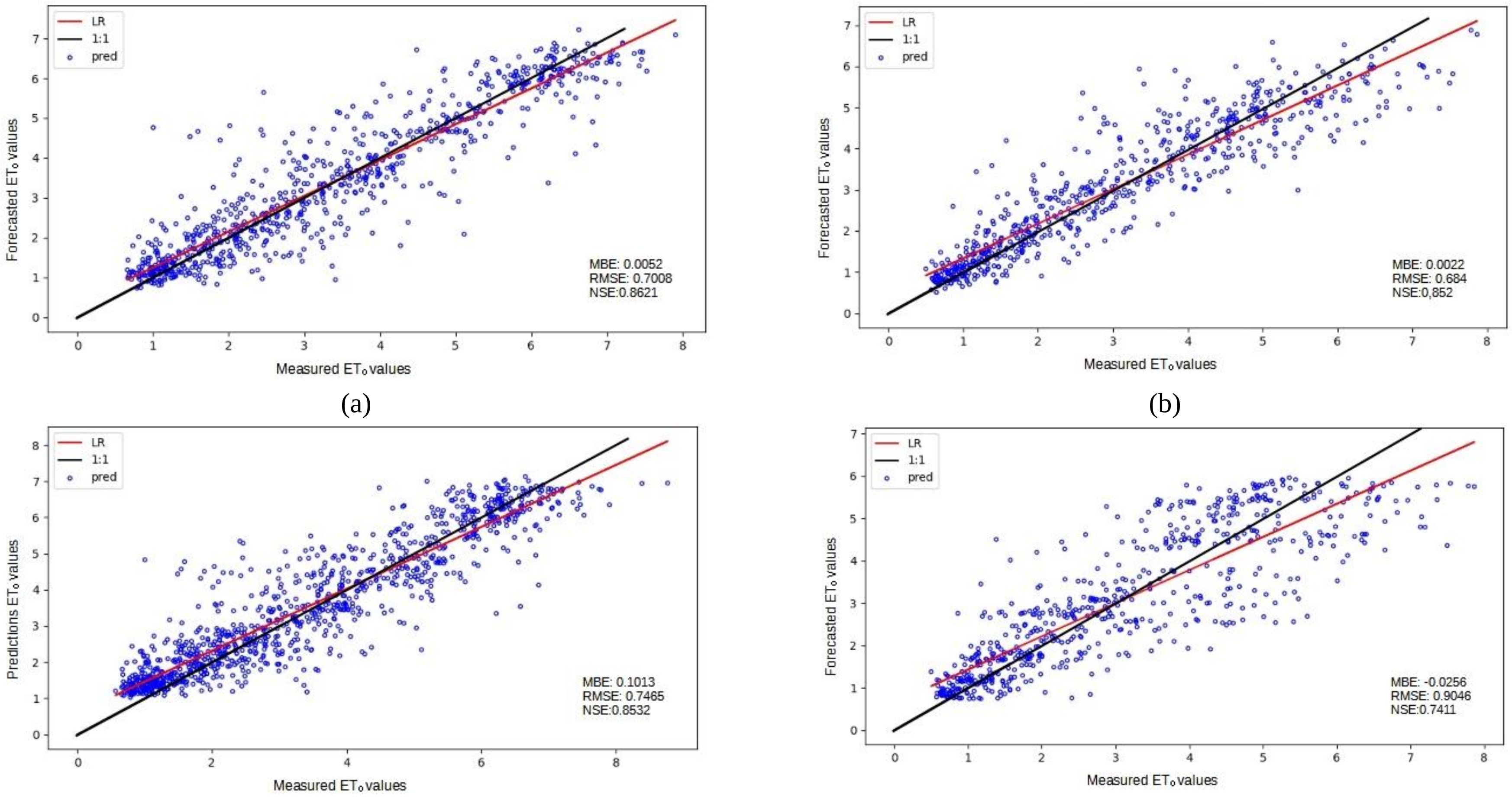
Conf.	TAB	CON	COR	ARO	MAG	Mean
I	0.806 (0.704)	0.886 (0.695)	0.720 (0.614)	**0.686** (0.605)	0.724 (0.648)	**0.764**
II	0.801 (0.709)	0.909 (0.697)	0.718 (0.618)	0.703 (0.615)	0.732 (**0.631**)	0.772
III	**0.786** (0.701)	0.920 (0.694)	**0.710** (0.633)	0.693 (0.603)	0.730 (0.643)	0.767

IV	0.794 (0.703)	0.897 (0.694)	0.724 (0.630)	0.693 (0.604)	0.734 (0.646)	0.768
V	0.812 (0.706)	0.914 (0.700)	0.741 (0.621)	0.704 (0.598)	0.732 (0.632)	0.780
VI	0.812 (0.709)	0.870 (**0.687**)	0.720 (0.622)	0.725 (0.602)	0.743 (0.645)	0.774
VII	0.805 (0.703)	0.902 (0.689)	0.728 (0.621)	0.710 (0.601)	0.733 (0.648)	0.775
VIII	0.805 (0.709)	0.925 (0.693)	0.737 (0.617)	0.717 (0.606)	0.725 (0.642)	0.781
IX	0.799 (0.708)	0.883 (0.694)	0.735 (0.642)	0.693 (0.613)	0.726 (0.639)	0.767
X	0.803 (0.704)	0.897 (0.699)	0.734 (0.620)	0.687 (0.613)	0.730 (0.641)	0.770
XI	0.811 (0.709)	0.931 (0.698)	0.740 (0.617)	**0.686** (**0.597**)	**0.702** (0.640)	0.774
XII	0.823 (0.712)	0.926 (0.697)	0.732 (0.640)	0.706 (0.605)	0.722 (0.641)	0.781
XIII	0.814 (0.708)	0.933 (0.691)	0.734 (**0.605**)	0.726 (0.615)	0.737 (0.642)	0.788
XIV	0.809 (0.714)	0.892 (0.688)	0.737 (0.643)	0.721 (0.615)	0.741 (0.643)	0.780
XV	0.811 (0.708)	0.899 (0.715)	0.730 (0.614)	0.698 (0.612)	0.721 (0.645)	0.771
XVI	0.824 (0.709)	0.904 (0.693)	0.722 (0.619)	0.706 (0.599)	0.736 (0.633)	0.778
XVII	0.810 (0.708)	0.921 (0.691)	0.753 (0.615)	0.726 (0.599)	0.734 (0.633)	0.788
XVIII	0.805 (0.707)	0.904 (0.718)	0.729 (0.622)	0.719 (0.606)	0.735 (0.647)	0.778
XIX	0.803 (0.707)	0.905 (0.688)	0.736 (0.616)	0.711 (0.605)	0.722 (0.633)	0.775
XX	0.816 (0.713)	0.879 (0.695)	0.733 (0.610)	0.719 (0.604)	0.747 (0.642)	0.778
XXI	0.801 (**0.700**)	0.920 (0.721)	0.725 (0.623)	0.696 (0.608)	0.738 (0.643)	0.776
XXII	0.792 (0.709)	0.893 (0.698)	0.728 (0.615)	0.709 (0.609)	0.722 (0.637)	0.768
XXIII	0.803 (0.713)	0.904 (0.696)	0.719 (0.627)	0.705 (0.604)	0.786 (0.643)	0.783
XXIV	0.823 (0.709)	0.917 (0.695)	0.741 (0.640)	0.696 (0.608)	0.731 (0.635)	0.781
XXV	0.821 (0.711)	**0.863** (0.691)	0.720 (0.618)	0.714 (0.613)	0.733 (0.655)	0.770

XXVI	0.822 (0.713)	0.894 (0.684)	0.736 (0.615)	0.711 (0.605)	0.730 (0.647)	0.778
XXVII	0.803 (0.710)	0.917 (0.699)	0.714 (0.627)	0.718 (0.612)	0.734 (0.636)	0.777

3.4. Overall discussion

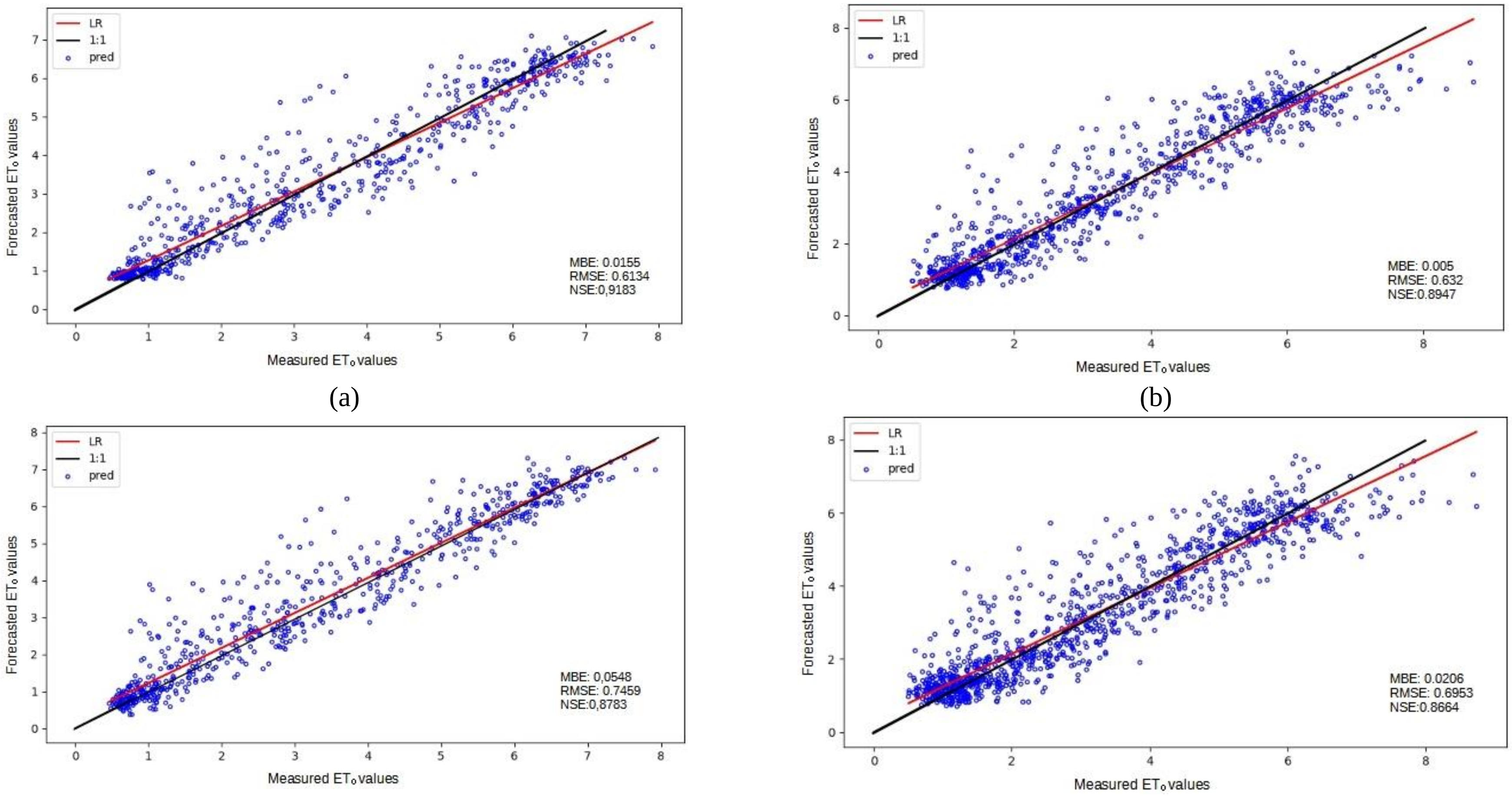
In this work, several aspects were evaluated in forecasting daily ET_0 at five locations from the Andalusia region (Southern Spain) with different geo-climatic conditions. Firstly, a new state-of-the-art architecture for NLP problems was assessed to forecast daily ET_0, the Transformers. Specifically, two different approaches were evaluated, TransformerCNN and TransformerLSTM, and they were compared to standard machine learning models like MLP, SVM, RF, or CNN, among others. In general, the results obtained using standard machine learning approaches such as RF, SVM, and ELM highly outperformed the rest of the models assessed in this work. Besides, Transformer-based models did not perform as expected in all cases when compared to standard ML models. However, their results were better than the baselines for most sites and cases (except for Conil). Secondly, another critical aspect to highlight in this work is that even using a self-attention mechanism (Transformer-based models), the use of 30 lag days instead of 15 lag days was not beneficial to forecasting daily ET_0. On the contrary, slightly better results were carried out when 15 lag days were used, and fewer bad outliers. Moreover, when comparing the different feature input configurations proposed in this study, none of them predominantly outperformed the rest, although configurations XIII, XIV, XX, and XXI were better on average. Figures 6.16, 6.17, and 6.18 show a scatter plot of measured vs. predicted ET_0 values using the best ML model and configuration for 1 and 7 days ahead.



(a) (b)

(c) (d)

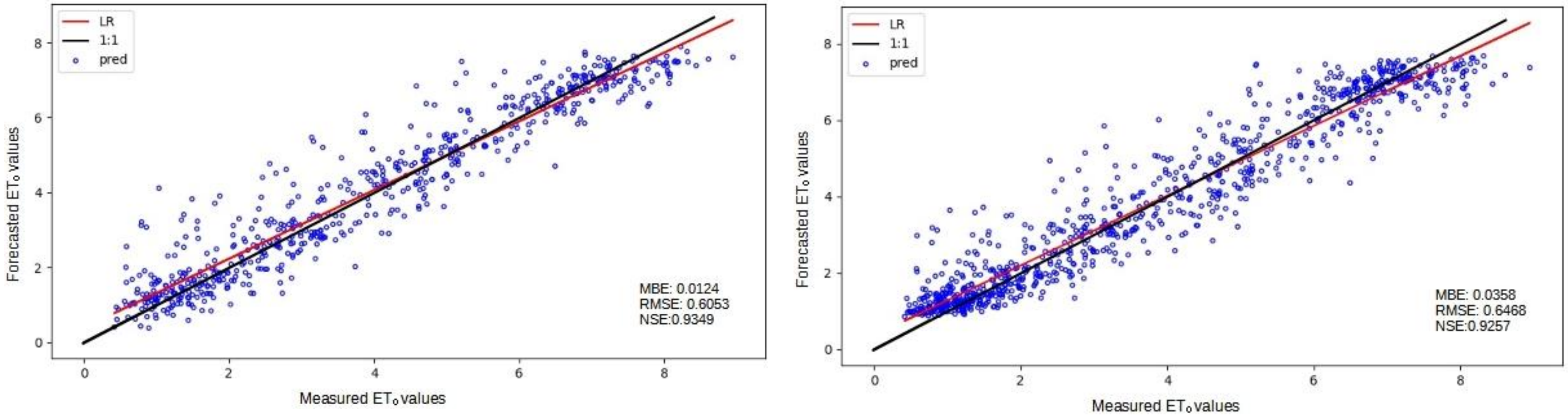
Figure 6. 16. Scatter plot for measured vs. predicted values for (a) forecast horizon 1 in Tabernas, (b) forecast horizon 1 in Conil de la Frontera, (c) forecast horizon 7 in Tabernas and (d) forecast horizon 7 in Conil de la Frontera.



(a) (b)

(c) (d)

Figure 6. 17. Scatter plot for measured vs. predicted values for (a) forecast horizon 1 in Aroche, (b) forecast horizon 1 in Málaga, (c) forecast horizon 7 in Aroche and (d) forecast horizon 7 in Málaga.



(a) (b)

Figure 6. 18. Scatter plot for measured vs. predicted values for for (a) forecast horizon 1 in Córdoba, (b) forecast horizon 7 in Córdoba

Furthermore, the results of the proposed models were significantly better than those reported by Ferreira and da Cunha [18] in terms of RMSE and NSE using different deep learning approaches in Brazil in AWS with an aridity index ranging from 0.3 to 1.6. The best NSE performances in Brazil ranged from 0.35 to 0.62 (approximately), whereas in this work, the best NSE values ranged from 0.60 to 0.95 (approximately). Besides, this work also obtained slightly better NSE values than those reported by Nourani et al. [17][17] using ensemble modeling in different weather stations from Iran, Turkey, and Cyprus. These previous works used temperature, relative humidity, solar radiation, and wind speed values as input features, whereas all the configurations of this work were temperature-based variables. Additionally, comparing the results to those obtained by de Oliveira e Lucas et al. [58][58], the assessed models in the present work outperformed their CNN and ensemble CNN results in Brazil.

In all, the models developed in this work, especially SVM, ELM, and RF are able to accurately forecast ET_0 for one week ahead using only temperature-based parameters and ET_0 past values. This issue is vital for improving crop irrigation scheduling, allowing adequate and anticipated planning, and contributing to agricultural production. Furthermore, providing reliable ET_0 future values positively impacts the current challenge of optimizing water resources management, especially in arid and semiarid locations.

4. Conclusions

In this work, several machine learning models have been developed and assessed for daily ET_0 forecasting from 1 to 7 days ahead using different input configurations, as well as different lag days. In general, all the ML approaches outperformed the baselines for all the forecast horizons and most locations, but SVM, RF, and ELM highly outperformed the rest of the models evaluated for most sites except for Conil de la Frontera, with unusual low wind speed values in its region. On the other hand, the Transformers were, on average, at the bottom of the ranking. Besides, all configurations obtained very similar results in terms of RMSE, but configurations I, III, IV and IX slightly outperformed the rest. The NSE values were above 0.85 for Conil, Tabernas, and Málaga and above 0.9 for Córdoba and Aroche for their best modeling. In terms of RMSE, the average performance for Tabernas was 0.92 mm/day, 1.00 mm/day for Conil, 0.81 mm/day for Córdoba, 0.80 mm/day for Aroche, and 0.78 mm/day for Málaga. It denotes a relationship in performance regarding the aridity index and the distance to the sea. For inland locations, the higher the aridity index, the less the error of forecasting ET_0 will be. On the other hand, for coastal sites, the higher the aridity index, the higher the error. Regarding MBE, most stations and models obtained very accurate values on average for most cases, with a mean performance value of 0.011 mm/day.

Further studies can deeply explore using these models in new regions with different geoclimatic conditions, different scenarios (a different time interval and a regional scenario), and for other parameters such as solar radiation or precipitation. Besides, an accurate feature selection or reduction could be researched because, as it could be stated based on the arose results, the configurations containing the worst related features based on Pearson correlation (HTx, HTn, HSr-HTn) obtained very accurate minimum and mean RMSE (Table 6.10 and Figure 6. 2). The approaches proposed in this work may result in greater efficiency for optimizing water resources, improving irrigation scheduling and anticipating the decision-making for agricultural goals. Finally, the creation of an open-source repository will allow novel scientists to apply these models using their own datasets, as well as experienced scientists to commit improvements with new features and architectures. Overall, the ultimate aim is to democratize the use of machine learning to more efficiently solve today's agricultural problems.

5. References

[1]. FAO The State of Food Security and Nutrition in the World 2021. The State of Food Security and Nutrition in the World 2021 2021, doi:10.4060/CB4474EN.

[2]. Allen, R.; Pereira, L.; Smith, M. Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-FAO Irrigation and Drainage; 1998; Vol. 56;.

[3]. Kwon, H.; Choi, M. Error Assessment of Climate Variables for FAO-56 Reference Evapotranspiration. Meteorology and Atmospheric Physics 2011, 112, 81–90, doi:10.1007/s00703-011-0132-1.

[4]. Estévez, J.; García-Marín, A.P.; Morábito, J.A.; Cavagnaro, M. Quality assurance procedures for validating meteorological input variables of reference evapotranspiration in mendoza province (Argentina). Agricultural Water Management 2016, 172, 96–109, doi:10.1016/j.agwat.2016.04.019.

[5]. Jabloun, M.; Sahli, A. Evaluation of FAO-56 Methodology for Estimating Reference Evapotranspiration Using Limited Climatic Data. Application to Tunisia. Agricultural Water Management 2008, 95, 707–715, doi:10.1016/j.agwat.2008.01.009.

[6]. Estévez, J.; P.Gavilán; J.V.Giráldez Guidelines on Validation Procedures for Meteorological Data from Automatic Weather Stations. Journal of Hydrology 2011, 402, 144–154, doi:10.1016/j.jhydrol.2011.02.031.

[7]. Estévez, J.; Padilla, F.L.; Gavilán, P. Evaluation and Regional Calibration of Solar Radiation Prediction Models in Southern Spain. Journal of Irrigation and Drainage Engineering 2012, 138, 868–879, doi:10.1061/(ASCE)IR.1943-4774.0000493.

[8]. WMO, W.M.O. Guide to Instruments and Methods of Observations; WMO: Geneva, Switzerland., 2018; Vol. 8; ISBN 978-92-63-10008-5.

[9]. George H. Hargreaves; Zohrab A. Samani Reference Crop Evapotranspiration from Temperature. Applied Engineering in Agriculture 1985, 1, 96–99, doi:10.13031/2013.26773.

[10]. Raziei, T.; Pereira, L.S. Estimation of ETo with Hargreaves-Samani and FAO-PM Temperature Methods for a Wide Range of Climates in Iran. Agricultural Water Management 2013, 121, 1–18, doi:10.1016/j.agwat.2012.12.019.

[11]. Ravazzani, G.; Corbari, C.; Morella, S.; Gianoli, P.; Mancini, M. Modified Hargreaves-Samani Equation for the Assessment of Reference Evapotranspiration in Alpine River Basins. Journal of Irrigation and Drainage Engineering 2012, 138, 592–599, doi:10.1061/(asce)ir.1943-4774.0000453.

[12]. Luo, Y.; Chang, X.; Peng, S.; Khan, S.; Wang, W.; Zheng, Q.; Cai, X. Short-Term Forecasting of Daily Reference Evapotranspiration Using the Hargreaves-Samani Model and Temperature Forecasts. Agricultural Water Management 2014, 136, 42–51, doi:10.1016/j.agwat.2014.01.006.

[13]. Karimi, S.; Shiri, J.; Marti, P. Supplanting Missing Climatic Inputs in Classical and Random Forest Models for Estimating Reference Evapotranspiration in Humid Coastal Areas of Iran. Computers and Electronics in Agriculture 2020, 176, 105633, doi:10.1016/j.compag.2020.105633.

[14]. Ferreira, L.B.; da Cunha, F.F. New Approach to Estimate Daily Reference Evapotranspiration Based on Hourly Temperature and Relative Humidity Using Machine Learning and Deep Learning. Agricultural Water Management 2020, 234, 106113, doi:https://doi.org/10.1016/j.agwat.2020.106113.

[15]. Yan, S.; Wu, L.; Fan, J.; Zhang, F.; Zou, Y.; Wu, Y. A Novel Hybrid WOA-XGB Model for Estimating Daily Reference Evapotranspiration Using Local and External Meteorological Data: Applications in Arid and Humid Regions of China. Agricultural Water Management 2021, 244, 106594, doi:10.1016/j.agwat.2020.106594.

[16]. Wu, L.; Peng, Y.; Fan, J.; Wang, Y.; Huang, G. A Novel Kernel Extreme Learning Machine Model Coupled with K-Means Clustering and Firefly Algorithm for Estimating Monthly Reference Evapotranspiration in Parallel Computation. Agricultural Water Management 2021, 245, 106624, doi:10.1016/j.agwat.2020.106624.

[17]. Nourani, V.; Elkiran, G.; Abdullahi, J. Multi-Step Ahead Modeling of Reference Evapotranspiration Using a Multi-Model Approach. Journal of Hydrology 2020, 581, 124434, doi:10.1016/j.jhydrol.2019.124434.

[18]. Ferreira, L.B.; da Cunha, F.F. Multi-Step Ahead Forecasting of Daily Reference Evapotranspiration Using Deep Learning. Computers and Electronics in Agriculture 2020, 234, 106113, doi:https://doi.org/10.1016/j.agwat.2020.106113.

[19]. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems 2017, 2017-December, 5999–6009.

[20]. Wu, S.; Xiao, X.; Ding, Q.; Zhao, P.; Wei, Y.; Huang, J. Adversarial Sparse Transformer for Time Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems; 2020; Vol. 2020-Decem.

[21]. Wu, N.; Green, B.; Ben, X.; O’Banion, S. Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. 2020.

[22]. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on

Time Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems; 2019; Vol. 32.

[23]. UNEP, N.M.; Thomas London, D. %J E.A.; Thomas London, D. World Atlas of Desertification. Edward Arnold 1992, 15–45.

[24]. Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. New Machine Learning Approaches to Improve Reference Evapotranspiration Estimates Using Intra-Daily Temperature-Based Variables in a Semi-Arid Region of Spain. Agricultural Water Management 2020, 245, 106558, doi:10.1016/j.agwat.2020.106558.

[25]. Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. Assessing Neural Network Approaches for Solar Radiation Estimates Using Limited Climatic Data in the Mediterranean Sea. In Proceedings of the The 3rd International Electronic Conference on Atmospheric Sciences (ECAS 2020); 2020.

[26]. Bellido-Jiménez, J.A.; Estévez Gualda, J.; García-Marín, A.P. Assessing New Intra-Daily Temperature-Based Machine Learning Models to Outperform Solar Radiation Predictions in Different Conditions. Applied Energy 2021, 298, 117211, doi:10.1016/J.APENERGY.2021.117211.

[27]. Estévez, J.; Gavilán, P.; García-Marín, A.P. Spatial Regression Test for Ensuring Temperature Data Quality in Southern Spain. Theoretical and Applied Climatology 2018, 131, 309–318, doi:10.1007/s00704-016-1982-8.

[28]. Islam, A.R.M.T.; Shen, S.; Yang, S.; Hu, Z.; Chu, R. Assessing Recent Impacts of Climate Change on Design Water Requirement of Boro Rice Season in Bangladesh. Theoretical and Applied Climatology 2019, 138, 97–113, doi:https://doi.org/10.1007/s00704-019-02818-8.

[29]. Yi, Z.; Zhao, H.; Jiang, Y. Continuous Daily Evapotranspiration Estimation at the Field-Scale over Heterogeneous Agricultural Areas by Fusing Aster and Modis Data. Remote Sensing 2018, 10, 1694, doi:https://doi.org/10.3390/rs10111694.

[30]. Sattari, M.T.; Apaydin, H.; Band, S.S.; Mosavi, A.; Prasad, R. Comparative Analysis of Kernel-Based versus ANN and Deep Learning Methods in Monthly Reference Evapotranspiration Estimation. Hydrology and Earth System Sciences 2021, 25, 603–618, doi:10.5194/hess-25-603-2021.

[31]. Tikhamarine, Y.; Malik, A.; Souag-Gamane, D.; Kisi, O. Artificial Intelligence Models versus Empirical Equations for Modeling Monthly Reference Evapotranspiration. Environmental Science and Pollution Research 2020, 27, 30001–30019, doi:10.1007/s11356-020-08792-3.

[32]. Huang, G. Bin; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: Theory and Applications. Neurocomputing 2006, 70, 489–501, doi:10.1016/j.neucom.2005.12.126.

[33]. Zhu, B.; Feng, Y.; Gong, D.; Jiang, S.; Zhao, L.; Cui, N. Hybrid Particle Swarm Optimization with Extreme Learning Machine for Daily Reference Evapotranspiration Prediction from Limited Climatic Data. Computers and Electronics in Agriculture 2020, 173, 105430, doi:10.1016/j.compag.2020.105430.

[34]. Akusok, A.; Björk, K.-M.; Miche, Y.; Lendasse, A. High Performance Extreme Learning Machines: A Complete Toolbox for Big Data Applications. Access, IEEE 2015, PP, 1, doi:https://doi.org/10.1109/ACCESS.2015.2450498.

[35]. Smola, A.J.; Schölkopf, B. A Tutorial on Support Vector Regression. Statistics and Computing 2004, 14, 199–222.

[36]. Chen, Z.; Zhu, Z.; Jiang, H.; Sun, S. Estimating Daily Reference Evapotranspiration Based on Limited Meteorological Data Using Deep Learning and Classical Machine Learning Methods. Journal of Hydrology 2020, 591, 125286, doi:10.1016/j.jhydrol.2020.125286.

[37]. de Oliveira, R.G.; Valle Júnior, L.C.G.; da Silva, J.B.; Espíndola, D.A.L.F.; Lopes, R.D.; Nogueira, J.S.; Curado, L.F.A.; Rodrigues, T.R. Temporal Trend Changes in Reference Evapotranspiration Contrasting Different Land Uses in Southern Amazon Basin. Agricultural Water Management 2021, 250, 106815, doi:10.1016/j.agwat.2021.106815.

[38]. Ghimire, S.; Deo, R.C.; Raj, N.; Mi, J. Deep Solar Radiation Forecasting with Convolutional Neural Network and Long Short-Term Memory Network Algorithms. Applied Energy 2019, 253, 113541, doi:10.1016/j.apenergy.2019.113541.

[39]. Kim, S.; Hong, S.; Joh, M.; Song, S.K. DeepRain: ConvLSTM Network for Precipitation Prediction Using Multichannel Radar Data. arXiv 2017.

[40]. Aloysius, N.; Geetha, M. A Review on Deep Convolutional Neural Networks. In Proceedings of the Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017; Institute of Electrical and Electronics Engineers Inc., February 7 2018; Vol. 2018-Janua, pp. 588–592.

[41]. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Computation 1997, 9, 1735–1780, doi:10.1162/neco.1997.9.8.1735.

[42]. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems 2017, 2017-December, 5999–6009.

[43]. Song, H.; Rajan, D.; Thiagarajan, J.J.; Spanias, A. Attend and Diagnose: Clinical Time Series Analysis Using Attention Models.

[44]. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems 2017, 2017-Decem, 5999–6009.

[45]. Wu, N.; Green, B.; Ben, X.; O’Banion, S. Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. 2020.

[46]. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing.; Association for Computational Linguistics (ACL), November 29 2020; pp. 38–45.

[47]. Mohammdi Farsani, R.; Pazouki, E. A Transformer Self-Attention Model for Time Series Forecasting. Journal of Electrical and Computer Engineering Innovations (JECEI) 2021, 9, 1–10, doi:10.22061/JECEI.2020.7426.391.

[48]. Chen, Z.; Zhu, Z.; Jiang, H.; Sun, S. Estimating Daily Reference Evapotranspiration Based on Limited Meteorological Data Using Deep Learning and Classical Machine Learning Methods. Journal of Hydrology 2020, 591, 125286, doi:https://doi.org/10.1016/j.jhydrol.2020.125286.

[49]. Alizamir, M.; Kisi, O.; Muhammad Adnan, R.; Kuriqi, A. Modelling Reference Evapotranspiration by Combining Neuro-Fuzzy and Evolutionary Strategies. Acta Geophysica 2020, 68, 1113–1126, doi:10.1007/s11600-020-00446-9.

[50]. Mohammadi, B.; Mehdizadeh, S. Modeling Daily Reference Evapotranspiration via a Novel Approach Based on Support Vector Regression Coupled with Whale Optimization Algorithm. Agricultural Water Management 2020, 237, 106145, doi:10.1016/j.agwat.2020.106145.

[51]. Gijsbers, P.; LeDell, E.; Thomas, J.; Poirier, S.; Bischl, B.; Vanschoren, J. An Open Source AutoML Benchmark. 2019.

[52]. Kotthoff, L.; Thornton, C.; Hoos, H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA. Journal of Machine Learning Research 2017, 18, 1–5.

[53]. Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019, 1946–1956.

[54]. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.T.; Blum, M.; Hutter, F. Auto-Sklearn:: Efficient and Robust Automated Machine Learning. In Proceedings of the Advances in Neural Information Processing Systems; 2015; Vol. 2015-Janua, pp. 2962–2970.

[55]. Hutter, F.; Kotthoff, L.; Vanschoren, J. Automated Machine Learning; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; The Springer Series on Challenges in Machine Learning; Springer International Publishing: Cham, 2019; ISBN 978-3-030-05317-8.

[56]. Borji, A.; Itti, L. Bayesian Optimization Explains Human Active Search. Advances in Neural Information Processing Systems 2013.

[57]. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE 2016, 104, 148–175, doi:10.1109/JPROC.2015.2494218.

[58]. de Oliveira e Lucas, P.; Alves, M.A.; de Lima e Silva, P.C.; Guimarães, F.G. Reference Evapotranspiration Time Series Forecasting with Ensemble of Convolutional Neural Networks. Computers and Electronics in Agriculture 2020, 177, 105700, doi:10.1016/j.compag.2020.105700.

Chapter 7

A regional machine learning method to outperform temperature-based reference evapotranspiration estimations in Southern Spain

Juan Antonio Bellido-Jiménez[1], Javier Estévez[1], Amanda P. García-Marín[1]

[1] Projects Engineering Area, Department of Rural Engineering, Civil Constructions and Projects Engineering, University of Córdoba, Spain

Abstract

Accurate estimations of reference evapotranspiration (ET_0) are crucial for determining crop water requirements and designing an adequate irrigation scheduling to optimize the use of water. In this work, a new clustering method to outperform the accuracy of ET_0 estimations only using temperature variables has been developed and assessed, based on the multifractal properties of the Diurnal Temperature Range (DTR). Thus, a more accurate weather stations' grouping method has been evaluated, regardless of their geographic location. All the datasets were collected from 89 automated weather stations in the period 2000-2018 and pooled into two main regions (1 and 2). In each region, an iterative procedure has been carried out: 1) selection of all the stations except the candidate one for the training procedure and 2) test procedure using the candidate station. The results showed that Machine Learning models (ML) highly outperformed Hargreaves-Samani (HS) in most of the cases, being Multilayer Perceptron (MLP) the most accurate over Extreme Learning Machine models (ELM). On average, the results obtained by MLP using the best configuration in the first region were better than those obtained in the second region. Specifically, the first region got an Root Mean Square Error (RMSE) = 0.6572 mm/d, Nash–Sutcliffe Efficiency (NSE) = 0.8967, Coefficient of Determination (R^2) = 0.9306 and Mean Bias Error (MBE) = |0.04|mm/d while the second region obtained an RMSE = 0.7034 mm/d, NSE = 0.8665, R^2 = 0.8968 and MBE = |0.045|mm/d. Regarding the seasonal performance, spring and autumn obtained the best NSE and R^2 results, whereas winter carried out the lowest RMSE values. This study

provides a new and more accurate methodology to improve ET_0 estimations on a regional basis and only using temperature data in the whole process.

1. Introduction

Reference Evapotranspiration (ET_0) quantifies the water needs of a reference surface, not short of water, only being affected by atmospheric conditions. Thereby, accurate ET_0 calculations are crucial to optimizing water resources management and developing a more precise irrigation scheduling. It is significantly impactful in semiarid and arid areas like Andalusia (Southern Spain), where precipitation is generally scarce [1], with very hot and dry summers and characterized by a structural water deficit [2]. The Food and Agriculture Organization of the United Nations (FAO) recommends the physic-based Penman-Monteith formula (FAO PM) as the sole standard method for its computation [3]. However, the number of required measured variables to compute FAO PM is very high (air temperature -T-, relative humidity -RH-, wind speed - u_2- and solar radiation -Rs-), which is translated into the costly installation and maintenance of the Automated Weather Stations (AWS). Besides, they usually contain non-reliable long-term datasets due to miscalibration and lack of maintenance issues [4,5]. Thus, the geographical density of complete AWS is generally low, mainly in developing countries and rural areas (Estévez et al., 2012; World Meteorological Organization, 2018).

In these terms, several studies have analyzed new methodologies to determine ET_0 using a limited dataset such as the well-known Hargreaves-Samani equation [8]. It calculates ET_0 using the daily maximum (Tx) and minimum air temperature (Tn), as well as the extraterrestrial solar radiation (Ra), which is calculated using the latitude of the site and the day of the year. Moreover, Machine Learning (ML) and Deep Learning (DL) models have gained extraordinary attention for achieving state-of-art performance to compute ET_0 in different climatic and aridity conditions. For example, Bellido-Jiménez et al. [9] analyzed the performance of Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM), Extreme Learning Machine (ELM), Generalized Regression Neural Network (GRNN) and Extreme Gradient Boosting (XGBoost) models using novel temperature-based variables such as $Energy_T$ and HTx in different sites of Andalusia (South of Spain). Despite no model significantly outperformed the rest, ELM and MLP usually gave the best results, whereas GRNN was significantly the worst in all cases and statistics. Besides, the introduction of these novel variables improved the ET_0 estimations compared to only using Tx, Tn, and Ra. Zhou et al. [10] evaluated Deep Factorization Machine (DeepFM), Gradient Boosting Decision Tree (GBDT), Light Gradient Boosting Method (LightGBM), Gradient Boosting with Categorial Feature Support (CatBoost), XGBoost, RF and SVM to estimate daily ET_0 in 12 sites in China. Muhammad et al. [11] studied the efficiency of Gene Expression Programming (GEP) to estimate daily ET_0 in 10 sites in Malaysia. They assessed up to 9 different configurations using air T, RH, Rs, and U_2 variables. In some cases, their GEP results could obtain similar results to FAO PM values. Adnan et al. [12] evaluated a new hybrid neuron fuzzy method based on Moth-flame optimization (MFO), Water Cycle optimization Algorithm (WCA), and Adaptive Neuro-Fuzzy Inference System (ANFIS). They tested up to 9 configurations using variables derived from T, sunshine duration (N), RH, and U_2. The use of ANFIS-WCMFO (a hybrid of both WCA and MFO) outperformed ANFIS-WCA and ANFIS-MFO. Sattari et al. [13] evaluated Gaussian Process Regression (GPR), Broyden-Fletcher-Goldfarb-Shanno Artificial Neural Network (BFGS-ANN), and Long Short-Term Memory (LSTM) to estimate monthly ET_0 using 10 different input features configurations (derived from Tm (mean air temperature), Tx, Tn and RH, U_2, and N) in Turkey. Bedi [14] evaluated the concept of Transfer learning (using the knowledge acquired from one task to improve others) to outperform ET_0 forecasts, addressing the problem of low efficiency due to small training datasets. Two approaches were considered: (1) training the models using datasets from the nearest cluster (different from the target) and eventually fine-tuning at the target cluster; (2) training the models using the nearest site to the target location within the same cluster and fine-tuning using information from the target site. Besides, Support Vector Regression (SVR), Deep-multilayer perceptron (DNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) were assessed using historical data to predict current timestamp values. The results showed slightly less prediction error and improved reliability when compared to a local training methodology. Maroufpoor et al. [15] evaluated the potential of MLP and Grey Wolf Optimization (GWO) to estimate ET_0 in Iran. Up to 7 input configurations were assessed, using Tx, Tn, RH, SSH, U_2, and P (precipitation), being the configuration containing Tx, Tn, and U_2 the most accurate model. Malik et al. [16] analyzed the performance of Multiple Model-Multilayer Perceptron (MM-MLP), Multivariate Adaptive Regression Spline (MARS), SVM, Multi-Gene Genetic Programming (MGGP), and M5Tree to estimate pan evapotranspiration (EPm) in two stations of India. MM-MLP and MGGP obtained the best results. Dimitriadou and Nikolakopoulos (2021) estimated annual Actual Evapotranspiration (Eta) using empirical methods such as Turc, modified Turc, and Coutagne and compared them to MODIS ET for the period 2016-2019 in Greece. The results arised that Turc could serve as an alternative to MODIS annual ET.

Despite the promising results of the aforementioned models, most of the developed models were trained and tested using data from the same station (local calibration), which is impractical in real-life applications (except for gap-filling). Consequently, it is needed to build new regional models that can be deployed in new sites (not used for calibration or for training the model). In these terms, Gavilán et al [18] evaluated a regional calibration based on the Diurnal Temperature Range (DTR) and annual wind speed of HS equation using 88 AWS in Andalusia (Southern Spain), covering a diversity of climatic conditions from semi-desert to subtropical coastal areas. Dai et al. [19] analyzed MLP and Multi Linear Regression (MLR) to estimate monthly ET_0 in three different regions of Mongolia according to its aridity index (arid, semi-arid, and sub-humid climate). Their models were trained using data from multiple stations within the same subregion (having similar climate conditions). Besides, they tested using daily T, RH, U_2, and N as inputs, obtaining RMSE=0.130 mm d^{-1} and R^2=0.986. Ramírez-Cuesta et al. [20] evaluated different approaches to estimate ET_0 at a regional scale using remote sensing and AWS data. They assessed two main approaches in order to calculate ET_0: using data from the closest station or the most similar weather station in Andalusia. Their results arose that the use of the most similar AWS resulted in an average error lower than 2.3%. However, their results were not homogeneous throughout the entire region, having high differences between the lower section of the Guadalquivir Valley (registering the smallest errors) and the mountainous areas (with abrupt changes in ET_0 values). Feng et al. [21] evaluated RF and GRNN to estimate ET_0 in two locations with similar climate conditions in the Sichuan basin (Southwest China). They performed cross-validation using data from both sites. Their results for the temperature-based approaches (using Tx, Tn, and Ra) were slightly more accurate in RF than in GRNN, but both models underestimated ET_0 since the actual values were greater at both stations. Ferreira et al. [22] assessed RF, MLP, and Convolutional Neural Networks (CNN) to estimate daily ET_0 in the state of Minas Gerais, in Brazil. They first analyzed using hourly temperature and relative humidity data on a regional and local scenario. The CNN approach exhibited the best results in all cases, providing performance gains of up to 11.9% in terms of Root Mean Squared Error (RMSE) and 7.8% in terms of Nash-Sutcliffe Efficiency (NSE). Regarding the regional scenario, they used 43 AWS with different climate conditions as training data, while 10 stations (representing the different climate variability from the training dataset) were used as testing data. Trajkovic et al. [23] provided an adjustment of the Thornthwaite equation for the Vojvodina region (Serbia)using data from 5 different AWS and introducing data from three meteorological stations outside the study area. This new approach (cTh) gave accurate estimates generally, being the best model in the site of Timisoara. Fang et al. [24] evaluated Partial Mutual Information (PMI), which captures the linear and nonlinear dependencies, to forecast ET_0 using MLR, SVM, and RF for 15 sites divided into two different regions (the Jing River Basin and the Beilvo River Basin) of China. Based on their results it could be stated that PMI presented a better performance for excluding redundant information than Partial-Correlation -based Input Selection (PCIS). Besides, SVM and RF could not always outperform the forecast of MLR models.Wu et al. [25] compared the results of Kernel Extreme Learning Machine (KELM), ANFIS, and RF, among others, to estimate monthly ET_0 on a regional scale using data from 26 stations in China. Finally, Ferreira et al. [26] evaluated different models (RF, MLP, LSTM, CNN, and LSTM-CNN) to forecast daily ET_0 for up to a week in 53 weather stations in Brazil following the procedure described in Ferreira et al. [22].

Generally, according to the aforementioned literature, several important methodological drawbacks need to be addressed in terms of developing, training, and deploying regional ET_0 models. On the one hand, due to the lack of open reliable meteorological datasets, several authors have grouped AWS data that belong to the same country or region (geographical criteria), and even places with no apparent geoclimatic relationship among them (places from different countries, latitudes, and climatic conditions). It has been stated by several authors the existence of different performances in ET_0 estimations according to the aridity index of the site, obtaining better modelings in humid zones and worse results in more arid and semi-arid locations [27,28]. Thus, developing a single model for sites with different geoclimatic conditions may arise in less efficient models. On the other hand, in order to deal with this issue, a possible solution could be making regions from sites that have similar aridity index values. Despite this last approach being convenient to obtain accurate models, the main problem is that it requires not only the same variables that FAO PM equation needs, but also precipitation. Thus, the development of robust and accurate regional models able to estimate ET_0 with limited-availability datasets has not been addressed yet. There is a great need to apply innovative approaches in non-used sites (training/calibration) with a very low budget but still having confident accuracy estimations. Thereby, in this work, a novel machine learning method has been assessed in order to outperform temperature-based regional ET_0 estimations. The stations' grouping criterion is based on a multifractal characterization of the Diurnal Temperature Range (DTR), following the work of Herrera-Grimaldi et al. [29], where 197 AWS from the Spanish Meteorological Agency (AEMET) that had similar multifractal behavior were grouped as the same cluster. These geographical clusters were taken into account using new AWSs from a different AWS network in Andalusia. The use of this multifractal characterization regionalization technique allows more accurate stations grouping regardless of their geographical distribution, only using daily temperature values for this purpose. It is worth noting that no similar work has previously used a clustering method based on DTR multifractal properties to outperform ET_0 estimations in a regional scenario.

Besides, due to the promising results of MLP and ELM for estimating and predicting ET_0 and solar radiation in this same region [9,30,31], outperforming the rest of the ML models significantly, both models are assessed in this work. Thus, the main objectives of this work are I) the evaluation of a clustering procedure based on the multifractal characterization of DTR, II) the assessment of MLP and ELM models to estimate ET_0 only using temperature-based features on a regional scenario, III) the analysis of different input feature configurations for each region and IV) the study the seasonal performance for each region.

2. Materials and methods

2.1. Source of data and data management

The study area is the region of Andalusia, which is located in southwest Europe. This region occupies a total area of 87 268 km^2 and ranges from 1 to 7 ºW longitude and from 37 to 39ºN latitude. It has a very heterogeneous climate, from the Mediterranean to the Continental (mountainous, oceanic, tropical, and sub-desert). Moreover, the datasets used in this work belong to 89 AWS from the Agroclimatic Information Network of Andalusia (RIA). All these AWSs measure air temperature, relative humidity of the air, wind speed and direction at 2 m height, solar radiation, and precipitation (the sensors' specifications are shown in Table 7.1) from April 2000 to July 2018 for a daily and sub-daily (every 30 min) sampling rate. They can be freely downloaded at https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web/ (Accessed 8th March 2022).

Table 7. 1. Sensors' specifications (Estévez et al., 2011)

Parameter	Sensor	Range	Accuracy
Wind speed	Young 05103	0 to 60 $m\ s^{-1}$	±0.3 $m\ s^{-1}$
Wind direction		0 to 360°	±3 °

Temperature	PT1000	-39.2 to 60 °C	±0.3 °C
Relative Humidity	Humicap 180	0.8 to 100 %	±3 % (90-100%) ±2 % (0.8-90%)
Precipitation	ARG 100		0.2 mm/tip
Solar radiation	Skye SP1110	350 to 1100 nm	±5 %

One of the key aspects for obtaining accurate modeling to estimate ET_0 on a regional basis is how we group the datasets of all the studied AWS. In this case, two main regions were considered based on the previous work of Herrera-Grimaldi et al [29], where a multifractal characterization of DTR values was carried out for 258 AWS from the Spanish AEMET (Meteorology Statal Agency) and grouped them based on their multifractal characteristics. In this work, the authors studied the scaling of the statistical moments to detect the multifractal behavior using the scaling exponent function K(q) that satisfies Eq. 7. 1 [32]:

$$\langle \varepsilon_{\lambda}^{q} \rangle \approx \lambda^{K(q)} \quad \text{Eq. 7. 1}$$

where $\langle \varepsilon_{\lambda}^{q} \rangle$ corresponds to the average qth moment of the intensity of the process for a λ resolution. The K(q) function becomes linear for q-values higher than the critical ($q_{critical}$), which is defined as γ_x (Eq. 7. 2).

$$\gamma_x = \max\left(K'(q)\right) \quad \text{Eq. 7. 2}$$

where γ is the order of singularity and γ_x its maximum value [33]. The γ_x gives important information about the process analyzed. Lower values are related to extreme values, whereas higher gamma values are characteristic of the smoothest processes. For further information, the following works can be revised [29,34,35]

The whole studied area was divided into two regions. Region 1 is composed of stations with low γ_x ($\gamma_x < 0.07216$) while region 2 includes those stations with γ_x higher than 0.07216. Figure 7. 1 shows the spatial distribution of the AWS and the considered regions, and Table 7. 2 the characteristics of all the AWS studied in this work as well as the geographical location of both regions.

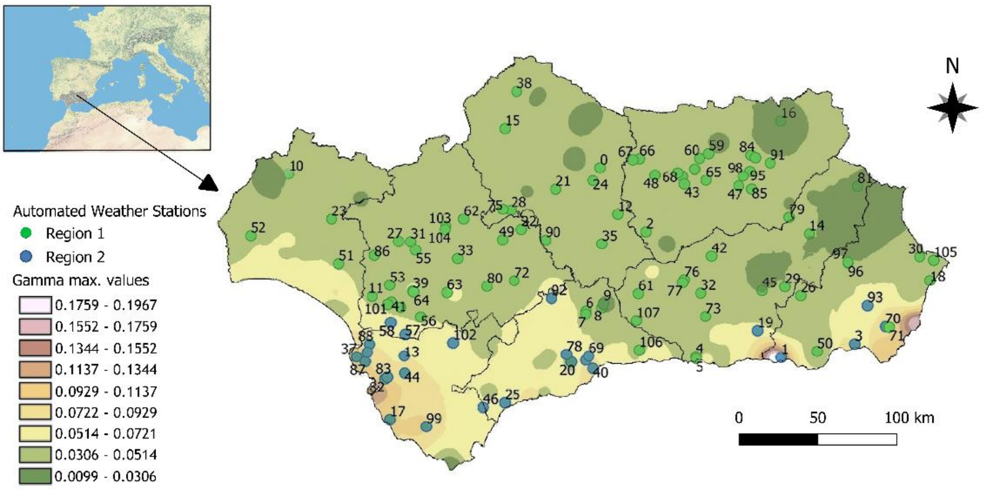


Figure 7. 1. Automated Weather Stations (AWS) in the Andalusia region and which region does each AWS belongs to RIA

Table 7. 2. Location and some characteristics of all the automated weather stations used in this work (Prov: Province; Lat: Latitude, Lon: Longitude; ALM: Almeria; CAD: Cadiz; COR: Cordoba; GRA: Granada; JAE: Jaen; HUE: Huelva; MAG: Malaga; SEV: Seville, u_2: Wind speed measured at two meters, Reg: Region).

ID	Name	Prov	Station code	Lat (◦N)	Lon (◦E)	Alt (m)	Mean u_2 (m/s)	Reg	Dataset length (available days)
0	Adamuz	Córdoba	COR02	38.00	-4,45	145	1.03	1	1999 - 2018 (6737)
1	Adra	Almería	ALM10	36.75	-2,99	2	1.55	2	2000 - 2018 (6447)
2	Alcaudete	Jaén	JAE06	37.58	-4,08	640	1.05	1	2001 - 2018 (6325)
3	Almería	Almería	ALM02	36.84	-2,40	5	1.43	2	2000 - 2018 (6525)
4	Almuñécar	Granada	GRA11	36.75	-3,68	29	0.968	1	2007 - 2018 (3935)
5	Antequera	Málaga	MAG03	37.06	-4,56	457	1.11	1	2000 - 2011 (3737)
6	Archidona	Málaga	MAG05	37.08	-4,43	516	0.92	1	2000 - 2012 (4258)
7	Aroche	Huelva	HUE06	37.96	-6,95	293	1.23	1	2000 - 2018 (6398)
8	Aznalcázar	Sevilla	SEV05	37.15	-6,27	2	2.22	1	2000 - 2018 (6308)
9	Baena	Córdoba	COR08	37.69	-4,31	310	1.19	1	2000 - 2018 (6411)
10	Basurta-Jerez de la Frontera	Cádiz	CAD01	36.76	-6,02	39	1.82	2	2000 - 2018 (6377)
11	Baza	Granada	GRA01	37.56	-2,77	718	1.82	1	2000 - 2018 (6531)
12	Bélmez	Córdoba	COR01	38.25	-5,21	503	2.10	1	2000 - 2018 (6585)
13	Chiclana de Segura	Jaén	JAE11	38.30	-3,00	571	1.41	1	2000 - 2018 (6387)
14	Conil de la Frontera	Cádiz	CAD05	36.33	-6,13	22	1.38	2	2000 - 2018 (5867)
15	Cuevas de Almanzora	Almería	ALM08	37.26	-1,80	28	1.00	1	1999 - 2018 (5972)
16	Cádiar	Granada	GRA07	36.92	-3,18	928	1.39	2	2000 - 2018 (6402)
17	Cártama	Málaga	MAG09	36.72	-4,68	78	0.85	2	2001 - 2018 (6172)
18	Córdoba	Córdoba	COR06	37.86	-4,80	94	1.63	1	2000 - 2018 (6396)
19	Ecija	Sevilla	SEV09	37.59	-5,08	109	1.55	1	2000 - 2018 (6352)
20	El Campillo	Huelva	HUE08	37.66	-6,60	383	1.71	1	2000 - 2018 (6411)
21	El Carpio	Córdoba	COR05	37.91	-4,50	171	1.70	1	2000 - 2018 (6305)

22	Estepona	Málaga	MAG04	36.44	-5,21	185	2.23	2	2000 - 2018 (6107)
23	Fiñana	Almería	ALM05	37.16	-2,84	958	2.5	1	2000 - 2018 (6686)
24	Guillena	Sevilla	SEV17	37.51	-6,06	48	1.52	1	2001 - 2018 (6148)
25	Hornachuelos	Córdoba	COR04	37.72	-5,16	130	1.31	1	2000 - 2018 (6415)
26	Huércal-Overa	Almería	ALM07	37.41	-1,88	303	1.35	1	1999 - 2018 (6572)
27	IFAPA Centro Las Torres-Tomejil	Sevilla	SEV19	37.51	-5,96	12	1.14	1	2005 - 2018 (4542)
28	IFAPA Centro Camino del Purchil	Granada	GRA101	37.17	-3,64	630	1.03	1	2005 - 2018 (4712)
29	IFAPA Centro Las Torres-Tomejil. Finca Tomejil	Sevilla	SEV101	37.40	-5,59	75	2.41	1	2001 - 2018 (6078)
30	IFAPA Centro Mengibar	Jaén	JAE104	37.94	-3,79	293	0.70	1	2008 - 2018 (3712)
31	IFAPA Centro de Cabra	Córdoba	COR101	37.50	-4,43	543	0.93	1	2003 - 2018 (5240)
32	IFAPA Centro de Campanillas	Málaga	MAG101	36.73	-4,56	63	0.75	2	2007 - 2018 (3921)
33	IFAPA Centro de Chipiona	Cádiz	CAD101	36.75	-6,40	7	1.99	2	2004 - 2018 (4940)
34	IFAPA Centro de Hinojosa del Duque	Córdoba	COR102	38.50	-5,12	543	1.36	1	2006 - 2018 (4328)
35	IFAPA Centro de Los Palacios	Sevilla	SEV21	37.19	-5,95	10	2.02	1	2007 - 2018 (3894)
36	IFAPA Churriana	Málaga	MAG07	36.67	-4,50	17	1.90	2	2001 - 2018 (6295)
37	Isla Mayor	Sevilla	SEV20	37.11	-6,12	2	2.33	1	2004 - 2018 (4972)
38	Iznalloz	Granada	GRA05	37.42	-3,55	921	1.88	1	2000 - 2018 (6498)
39	Jaén	Jaén	JAE15	37.89	-3,77	299	1.05	1	2001 - 2018 (6203)
40	Jerez de la Frontera	Cádiz	CAD02	36.64	-6,01	17	1.45	2	2000 - 2018 (6444)
41	Jerez del Marquesado	Granada	GRA06	37.19	-3,15	1,201	1.79	1	2000 - 2018 (6469)
42	Jimena de la Frontera	Cádiz	CAD07	36.41	-5,38	50	1.73	2	2000 - 2018 (6402)
43	Jódar	Jaén	JAE103	37.88	-3,33	486	0.75	1	2006 - 2018 (4431)
44	La Higuera de Arjona	Jaén	JAE12	37.95	-4,01	257	1.03	1	2001 - 2018 (6181)
45	La Luisiana	Sevilla	SEV10	37.53	-5,23	173	1.19	1	2000 - 2018 (6376)
46	La Mojonera	Almería	ALM01	36.79	-2,70	137	1.80	1	2000 - 2018 (6696)
47	La Palma del Condado	Huelva	HUE09	37.37	-6,54	171	1.56	1	2000 - 2018 (6367)

48	La Puebla de Guzmán	Huelva	HUE07	37.55	-7,25	248	1.91	1	2000 - 2018 (6306)
49	La Puebla del Río I	Sevilla	SEV07	37.23	-6,13	3	1.72	1	2000 - 2018 (6396)
50	La Puebla del Río II	Sevilla	SEV08	37.08	-6,05	2	1.89	1	2000 - 2018 (6316)
51	La Rinconada	Sevilla	SEV12	37.46	-5,92	25	1.86	1	2000 - 2018 (6286)
52	Las Cabezas de San Juan	Sevilla	SEV02	37.02	-5,88	13	2.12	1	2000 - 2018 (6412)
53	Lebrija	Sevilla	SEV03	36.98	-6,13	1	2.17	2	2000 - 2018 (6379)
54	Linares	Jaén	JAE09	38.06	-3,65	432	1.61	1	2000 - 2018 (6420)
55	Loja	Granada	GRA03	37.17	-4,14	463	1.90	1	2000 - 2018 (6476)
56	Lora del Río	Sevilla	SEV15	37.66	-5,54	44	1.38	1	2000 - 2018 (6430)
57	Los Molares	Sevilla	SEV16	37.18	-5,67	77	2.31	1	2001 - 2018 (6292)
58	Mancha Real	Jaén	JAE07	37.92	-3,60	407	1.92	1	2000 - 2018 (6477)
59	Marmolejo	Jaén	JAE10	38.06	-4,13	208	1.39	1	2000 - 2009 (3328)
60	Málaga	Málaga	MAG01	36.76	-4,54	55	1.33	2	2000 - 2018 (6422)
61	Níjar	Almería	ALM11	36.95	-2,16	169	1.85	2	2001 - 2018 (6046)
62	Osuna	Sevilla	SEV11	37.26	-5,13	198	2.18	1	2001 - 2018 (6046)
63	Padul	Granada	GRA10	37.02	-3,60	747	1.10	1	2001 - 2018 (6298)
64	Palma del Rio	Córdoba	COR09	37.73	-5,23	58	1.32	1	2007 - 2018 (4112)
65	Pinos Puente	Granada	GRA04	37.26	-3,77	594	1.80	1	2000 - 2014 (5001)
66	Pizarra	Málaga	MAG08	36.77	-4,72	71	1.54	2	2001 - 2018 (6278)
67	Pozo Alcón	Jaén	JAE02	37.67	-2,93	881	1.44	1	2000 - 2018 (6488)
68	Puebla Cazalla	Sevilla	SEV18	37.22	-5,35	193	1.33	1	2001 - 2018 (6239)
69	Puebla de Don Fadrique	Granada	GRA02	37.88	-2,38	1017	2.70	1	2000 - 2018 (6395)
70	Puerto de Santa María	Cádiz	CAD08	36.62	-6,15	20	2.55	2	2001 - 2010 (3486)
71	Sabiote	Jaén	JAE04	38.08	-3,24	791	2.26	1	2000 - 2018 (6446)
72	San José de los Propios	Jaén	JAE03	37.86	-3,23	494	2.05	1	2000 - 2018 (6478)
73	Sanlúcar La Mayor	Sevilla	SEV13	37.42	-6,26	63	1.40	1	2000 - 2018 (6122)

74	Santaella	Córdoba	COR07	37.52	-4,89	196	1.75	1	2000 - 2018 (6442)
75	Santo Tomé	Jaén	JAE14	38.03	-3,08	537	0.74	1	2001 - 2018 (6246)
76	Sierra Yeguas	Málaga	MAG06	37.14	-4,84	467	2.30	2	2000 - 2018 (6292)
77	Tabernas	Almería	ALM04	37.09	-2,30	502	1.91	2	2000 - 2018 (6693)
78	Torreblascopedro	Jaén	JAE05	37.99	-3,69	275	0.91	1	2000 - 2018 (6493)
79	Torreperogil	Jaén	JAE101	37.97	-3,24	535	1.97	1	2006 - 2018 (4166)
80	Tíjola	Almería	ALM12	37.38	-2,46	776	1.76	1	2001 - 2018 (6022)
81	Úbeda	Jaén	JAE08	37.94	-3,30	343	0.95	1	2000 - 2018 (6407)
82	Vejer de la Frontera	Cádiz	CAD06	36.29	-5,84	13	2.70	2	2000 - 2018 (6392)
83	Villacarrillo	Jaén	JAE102	38.06	-3,20	649	0.38	1	2006 - 2018 (4396)
84	Villamartín	Cádiz	CAD04	36.84	-5,62	146	1.81	2	2000 - 2018 (6201)
85	Villanueva del Río y Minas	Sevilla	SEV14	37.61	-5,68	38	1.08	1	2000 - 2010 (3312)
86	Virgen de Fátima-Cuevas de Almanzora	Almería	ALM06	37.39	-1,77	173	2.14	1	1999 - 2018 (6745)
87	Vélez-Málaga	Málaga	MAG02	36.80	-4,13	33	1.59	1	2001 - 2018 (6053)
88	Zafarraya	Granada	GRA08	36.99	-4,15	892	1.18	1	2000 - 2018 (6379)

In order to guarantee reliable results, an essential prerequisite is the application of quality assurance tests, where erroneous values are flagged and further evaluated. The guidelines described by Estévez et al [4,36] have been followed, applying range (tests regarding the specifications for each sensor as well as physical and climate extremes for each location and variable), internal consistency (tests based on physics and climatological consistency for each location and variable), step (tests based on time consistency), persistence (checking the variability of the measurements), and spatial consistency (designed to detect gross errors for observations) tests to all meteorological data. Afterward, the daily ET_0 was calculated using the FAO PM formula [3], which was used as the target variable, as well as the rest of the daily temperature-based features [9]: the maximum and minimum air temperature (Tx and Tn, respectively), DTR (Tx – Tn), the integral of the half-hourly temperature values of a day ($Energy_T$), the time of day in hours when the Tx and Tn occur (HTx and HTn, respectively). Besides, 70 different input configurations were previously assessed, where the following variables were considered: Tx (day-1), Tx (day), Tn (day-1), Tn (day), Tn (day+1), Ra (day), TD (day), HTx (day-1), HTx (day), HTn (day-1), HTn (day), $Energy_T$ (day-1) and $Energy_T$ (day). However, only the ten best configurations have been further discussed in this work (see Table 7. 3). It is worth noting that Tn (day+1) was analyzed based on the work of Bristow and Campbell [37], where they use Tn (day+1) to calculate Tm (average temperature).

Table 7. 3. All configurations and the features assessed in this work for each one. Ra (day) corresponds to Extraterrestrial Solar radiation from the same day, Tx (day-1) corresponds to maximum daily air temperature from the previous day, Tx (day) corresponds to the maximum daily air temperature from the same day, Tn (day-1) corresponds to minimum daily air temperature from the previous day, Tn (day) corresponds to minimum daily air temperature from the same day, DTR (day-1) corresponds to the difference between Tx and Tn from the previous day, DTR (day) corresponds to the difference between Tx and Tn from the same day, Energy_T (day-1) represents the integral of the half-hourly temperature records from the previous day, Energy_T (day-1) represents the integral of the half-hourly temperature records from the same day, HTx (day) represents the time when Tx occurs from the same day and HTn (day) represents the time when Tn occurs from the same day

Configuration	**Ra (day)**	**Tx (day-1)**	**Tx (day)**	**Tn (day-1)**	**Tn (day)**	**Tn (day+1)**	**DTR (day)**	**Energy_T (day-1)**	**Energy_T (day)**	**HTx (day)**	**HTn (day)**
I	X	X	X		X	X	X		X	X	
II	X		X	X	X		X		X		
III	X	X	X	X	X		X		X		
IV	X	X	X		X			X	X	X	
V	X		X		X				X		X
VI	X	X	X		X				X	X	
VII	X		X	X	X				X		
VIII	X	X	X		X				X		
IX	X	X	X		X	X			X		
X	X		X		X			X			

In order to achieve the goals of this work, the following control loop was followed: 1) All the available datasets from RIA were collected on a daily and intradaily (every half hour) basis; 2) several quality assurance tests (range, step, internal consistency, persistence, and spatial consistency tests) were applied following the procedures described in Estévez et al. (2016, 2011); 3) all the stations were grouped into two regions, based on the multifractal characterization carried out by Herrera-Grimaldi et al [29]; 4) then, to perform an exhaustive assessment in each region, one ML model was developed for each AWS (a form of cross-validation on a regional basis). At each site, the model was trained using the dataset from all AWS in the region except for the candidate station, while the dataset from the selected site was later used for testing purposes. For example, in the case of COR06 (Córdoba) which is part of the first region, the model to estimate ET_0 at this site was trained using the whole dataset of region 1 except the data from COR06; 5) Then, for each AWS in the region, and in order to tune the different hyperparameters, a hold-out technique is used instead, due to the enough size of the training dataset (almost 400 000 registers for the first region and more than 100 000 registers for the second) to avoid overfitting [38,39]. In these terms, the Bayesian optimization and validation dataset (30% of the training dataset) were used; 6) Finally, the final model was fully trained using the initial training dataset and assessed using the testing dataset (the data from the candidate AWS). Figure 2 shows a graphical summary of this methodology

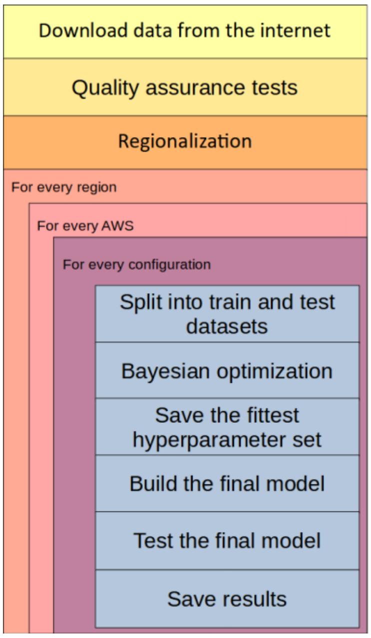


Figure 7. 2. Methodology

2.2. Reference evapotranspiration

In this work, the ET_0 values, calculated using the ET_0 equation [3], were used as the target values for the developed models. It can be mathematically expressed as Eq. 7. 3:

$$\mathrm{ET_0} = \frac{0.408\Delta(\mathrm{Rn} - G) + \gamma \frac{900}{Tm + 273} u_2(\mathrm{es} - \mathrm{ea})}{\Delta + \gamma(1 + 0.34\mathrm{u_2})} \qquad \text{Eq. 7. 3}$$

where ET_0 is the reference evapotranspiration (mm day^{-1}) calculated using this equation, 0.408 corresponds to a coefficient (MJ^{-1} m^2 mm), Δ is the slope of the saturation vapor pressure versus temperature curve (kPa $°C^{-1}$), Rn is the net radiation calculated at the crop surface (MJ m^{-2} day^{-1}), G is the soil heat flux density at the soil surface (MJ m^{-2} day^{-1}), γ is the psychrometric constant (kPa $°C^{-1}$), T is the mean daily air temperature (°C), u_2 is the mean daily wind speed at 2 m height (m s^{-1}), and es and ea are the saturation vapor pressure and the mean actual vapor pressure, respectively (kPa).

However, it is not always possible to find AWS collecting all the required parameters by FAO PM due to the expensive implementation and maintenance of these devices, and the existence of missing and non-reliable wind speed and solar radiation data is high [6,40]. Thus, different temperature-based approaches have been developed to handle this problem, such as the Hargreaves-Samani (HS) equation, developed by Hargreaves and Samani (1985) as an empirical equation that only uses Tx (day), Tm (day), Tn (day) and Ra (day) to compute ET_0 (Eq. 7. 4). HS has been widely applied to different climates and countries [27,41–43].

$$\mathrm{ET}_{0\ \mathrm{HS}} = 0.0135\mathrm{k}(\mathrm{Tm} + 17.8)\sqrt{Tx - Tn}\, Ra \qquad \text{Eq. 7. 4}$$

where $ET_{0\ HS}$ is the reference evapotranspiration (mm day^{-1}) calculated using HS equation and k is a coefficient (0.16 for inland and 0.19 for coastal locations). Additionally, even though HS is the most common worldwide temperature-based method there are other temperature-based methodologies, such as the Thornthwaite equation [44] or the approximation from FAO PM [3].

2.3. Machine learning models

All the models were developed using the public repository from GitHub AgroML [30], which can be publicly accessed at https://github.com/Smarity/AgroML (Accessed March 16, 2022). Besides, all the computations were carried out in a server workstation with the following characteristics: 1 CPU AMD Ryzen Threadripper PRO 3955WX 16-cores (3,9 GHz), 1 GPU Nvidia RTX 3090, and 128 GB of RAM. Based on the performance obtained in previous works [9], two kinds of models have been assessed: Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM).

2.3.1. Multilayer Perceptron

The multilayer perceptron (MLP) is based on the neurons from the biological nervous systems, which is one of the reasons for its high reputation. Besides, it is characterized by easy coding and promising results within ET_0 estimations in most cases [22,45–47]. They are structured into three different layers, the input layer that represents the inputs of the model, the hidden layer or layers where the neurons are located, and the output layer that represents the outputs of the model (ET_0 in this case). For example, for an MLP architecture with two inputs, one single hidden layer containing three neurons, and one output, the structure is represented by figures and equations 7. 5, 7. 6, 7. 7, and 7. 8.

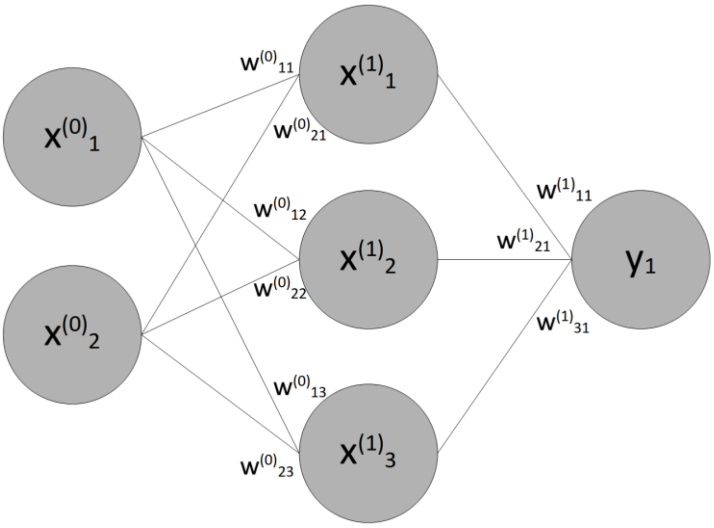


Figure 7. 3. Multilayer perceptron architecture for a model with two inputs ($x^{(0)}_1$ and $x^{(0)}_2$), one single hidden layer containing 3 neurons ($x^{(1)}_1$, $x^{(1)}_2$ and $x^{(1)}_3$) and one output (y1). W represents the weights for each connection.

Eq. 7. 5
$$x_1^{(1)} = f\left(x_1^{(0)}w_{11}^{(0)} + x_2^{(0)}w_{21}^{(0)} + b_1^{(0)}\right)$$

Eq. 7. 6
$$x_2^{(1)} = f\left(x_1^{(0)}w_{12}^{(0)} + x_2^{(0)}w_{22}^{(0)} + b_2^{(0)}\right)$$

Eq. 7. 7
$$x_3^{(1)} = f\left(x_1^{(0)}w_{12}^{(0)} + x_2^{(0)}w_{22}^{(0)} + b_3^{(0)}\right)$$

Eq. 7. 8
$$y_1 = f\left(x_1^{(1)}w_{11}^{(1)} + x_2^{(1)}w_{21}^{(1)} + x_3^{(1)}w_{31}^{(1)} + b_1^{(1)}\right)$$

where $x^{(0)}_1$ and $x^{(0)}_2$ represents the inputs of the model, $x^{(1)}_1$, $x^{(1)}_2$ and $x^{(1)}_3$ represents the neurons of the hidden layer, y_1 the output itself and w represents the weights of each connection. The weights of the model are automatically updated using a backpropagation algorithm during the training period.

2.3.2. Extreme Learning Machine

Extreme Learning Machine (ELM) models are a novel approach [48] composed of a single hidden layer feedforward neural network with much fewer computational requirements than MLPs. The main feature of this model is that the weights of the inputs are automatically generated, while the weights of the outputs are analytically calculated. Thereby, there is no need to carry out a backpropagation training process, which makes these models have a very low computational cost and high demand for RAM if the dataset is very large. Several works have assessed its use in ET_0 estimation and prediction worldwide [25,49].

2.4. Bayesian optimization

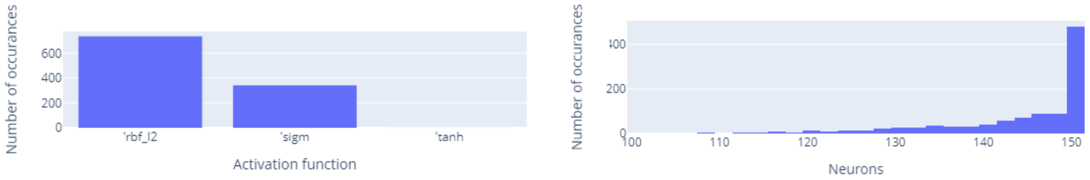
A critical aspect of obtaining accurate predictions in ML models is choosing a correct hyperparameter set. Depending on the chosen hyperparameters, the performance could dramatically change from accurate to very poor results. The hyperparameters of ML models modify their architecture and internal functional configuration, for example, the number of hidden layers, the hidden neurons within each layer, and the activation function, among others, in MLP. One of the most common techniques to deal with hyperparameter selection is a trial and error technique, where different values are evaluated and the best one is chosen. However, using this technique there is no assurance of the accurate performance of our model. A more efficient approach is the use of Bayesian Optimization (BO), based on the Bayes theorem and with high relation to human behavior when addressing this problem [50,51] because it takes into account the previous results to select the next set of hyperparameters to test. Even though there exist other algorithms that deal with hyperparameter tuning such as Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Genetic Algorithms (GA), Whale Optimization Algorithm (WOA), among others [52–54], the popularity of Bayesian Optimization is very high in automated machine learning libraries like Auto-Weka 2.0 [55], Auto-Keras [56] and Auto-Sklearn [57] and they can be further consulted in Hutter et al. (2019). Besides, its popularity in ET_0 models is high [59,60]

The Bayesian optimization can be explained using four steps: (1) The hyperparameter space is set, defining the limit values of all the hyperparameters; (2) the algorithm asses several random values for all the hyperparameters; (3) the algorithm set the next hyperparameter values to be evaluated taking into account the previous experience, weighing up between exploitation (choosing values known to have outstanding results) or exploration (choosing values with higher uncertainty) behavior; (4) if it has not finished yet, it repeats step 3.

In this work, the Bayesian Optimization algorithm from AgroML has been used, which had been implemented using the Scikit-Optimize library. Both models, ELM and MLP, were configured to carry out 50 Bayesian epochs (40 of them were randomly generated) and the Mean Absolute Error (MAE) as the objective function. Table 7. 4 shows the hyperparameter space of the models and Figures 4 and 5 show the BO results from ELM and MLP, respectively.

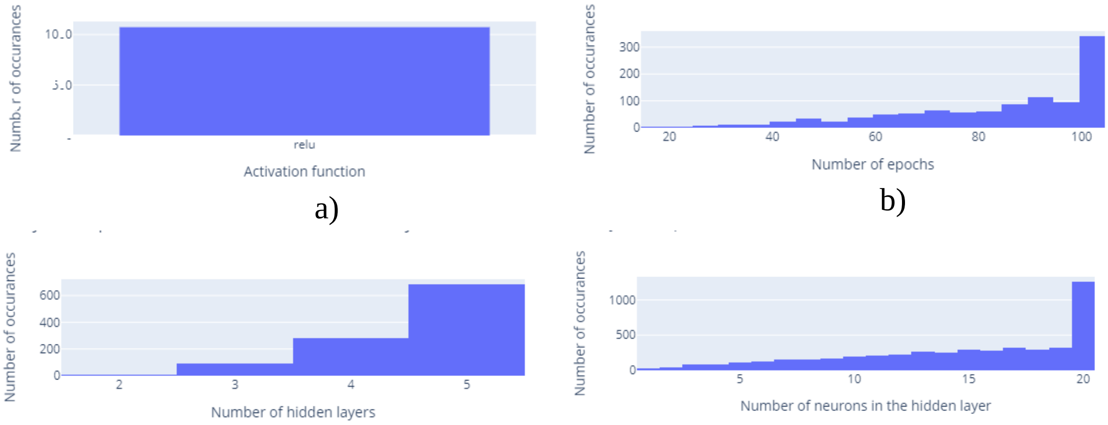
Table 7. 4. Hyperparameter space. MLP refers to Multilayer Perceptron, ELM refers to Extreme Learning Machine, relu refers to Rectified Linear Unit activation function, sigm to the sigmoid activation function, tanh to the hyperbolic tangent activation function, rbf_l2 a Radial Basis Function using L2 regularization

Model	Hyperparameter	Space
MLP	Number of hidden layers	Up to 5
	Number of neurons	Up to 20
	Activation function	relu, sigm and tanh
	Optimization function	Adam
	Maximum number of epochs	150
ELM	Maximum number of neurons	150
	Activation function	linear, sigm, tanh, rbf_l1, rbf_l2, rbf_Linf



a) b)

Figure 7. 4. BO results for the ELM hyperparameters, where a) represents the chosen activation function and b) the number of neurons in the hidden layer. Rbf_l2 represents a Radial Basis Function using L2 regularization, sigm represents a sigmoid activation function and tanh represents a hyperbolic tangent activation function



a) b)

c) d)

Figure 7. 5. BO results for the MLP hyperparameters, where a) represents the chosen activation function, b) the number of training epochs, c) the number of Hidden layers and d) the number of neurons on each hidden layer. Relu represents Rectified Linear Unit activation function

2.5. Statistical analysis

All the models were analyzed, in order to evaluate their performance, using the MBE, the RMSE, the R^2, NSE, and the Global Performance Indicator (GPI). The MBE, RMSE, R^2, NSE, and GPI can be defined as Equations 7. 9, 7.10, 7. 11, 7. 12, and 7. 13, respectively

$$MBE = \frac{1}{m}\sum_{i=1}^{m} meas_i - pred_i \qquad \text{Eq. 7. 9}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(meas_i - pred_i)^2} \qquad \text{Eq. 7. 10}$$

$$R^2 = \frac{\left(\sum_{i=1}^{m}(meas_i - \overline{meas})\left(pred_i - \overline{pred}\right)\right)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2\sum_{i=1}^{m}\left(pred_i - \overline{pred}\right)^2} \qquad \text{Eq. 7. 11}$$

$$NSE = 1 - \frac{\sum_{i=1}^{m}(meas_i - pred_i)^2}{\sum_{i=1}^{m}(meas_i - \overline{meas})^2} \qquad \text{Eq. 7. 12}$$

$$GPI_j = \sum_{j=1}^{4} \alpha_m (\overline{stat}_m - stat_{mj}) \quad \text{Eq. 7. 13}$$

where i represents each record, *m* represents the number of records, *x* represents the ET_0 values calculated using the FAO56 PM formula, *y* represents the estimations from models, α is a coefficient (being 1 for RMSE and -1 for R^2 and NSE), j represents each statistical performance indicator (MBE, RMSE, R^2 and NSE), j defines a specific configuration, $stat_{mj}$ represents the performance indicator value normalized (it is worth noting that instead of using the original MBE values, their absolute values were calculated), and the bar denotes the mean value. Besides, in order to compare how ML models performed compared to HS, the difference in the different statistics is shown as $STAT_{BestML\text{-}HS} = STAT_{BestML} - STAT_{HS}$, where STAT represents any of the statistics (MBE, RMSE, R^2, and NSE).

3. Results and discussion

To further evaluate the different results of this work, all the graphs can be interactively revised as an HTML file in the Additional files section.

3.1.The evaluation of ML models in the different regions

Figures 7. 6 and 7. 7 show the RMSE values for the best configuration at each site for regions 1 and 2, respectively. Generally, both ELM and MLP (ML) highly outperformed HS in most of the cases, obtaining the highest improvement in JAE08 - Úbeda ($RMSE_{BestML\text{-}HS}$ = -0.533 mm/d) and CAD07 - Jimena de la Frontera ($RMSE_{BestML\text{-}HS}$ = -0.410 mm/d) for regions 1 and 2, respectively. The lowest RMSE values were obtained in the stations HUE09 - La Palma del Condado ($RMSE_{BestML}$ = 0.473 mm/d) for the first region and in MAG01 - Málaga for the second ($RMSE_{BestML}$ = 0.548 mm/d). On the one hand, most of the sites got RMSE values below 0.8 mm/d for both regions, even below 0.6 mm/d for a high percentage of AWSs. On the other hand, very few locations (JAE102 - Villacarrillo, ALM06 – Virgen de Fátima-Cuevas de Almanzora , and SEV11 - Osuna) obtained RMSE values above 1.0 mm/d using ML models. Besides, only in 2 locations from the first region (GRA02 – Puebla de Don Fabrique and SEV101 – IFAPA Centro las Torres-Tomejil. Finca Tomejil), the ML underperformed HS with an error of around 0.1 mm/d. It is also worth noting that in general, the average ML performance in the first region was better than in the second ($RMSE_{reg1}$=0.665 ± 0.151 mm/d and $RMSE_{reg2}$=0.712 ± 0.113 mm/d) although in this last region MLP and ELM significantly improve the results of HS ($RMSE_{BestML\text{-}HS}$ = -0.1571 mm/d and $RMSE_{BestML\text{-}HS}$= -0.2123 mm/d for region 1 and 2, respectively)

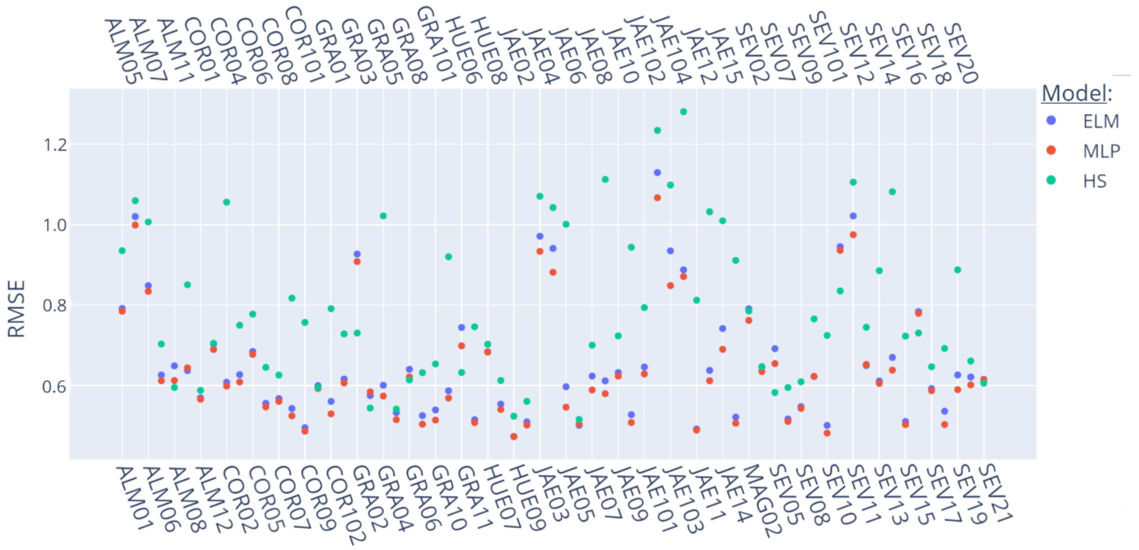


Figure 7. 6. RMSE values for the best configurations in region 1. See table 2 for further information about each location.

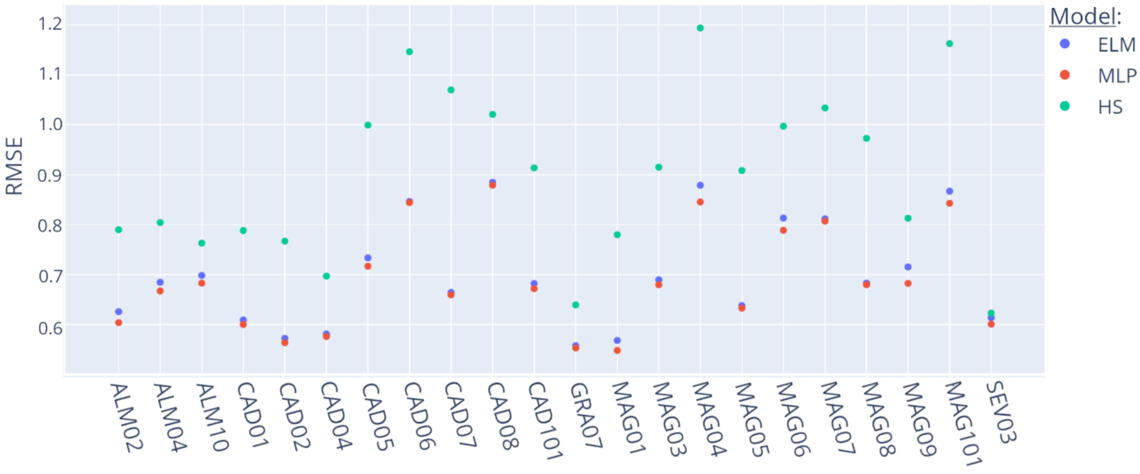


Figure 7. 7. RMSE values for the best configurations in region 2. See table 2 for further information about each location

Regarding the NSE values, Figures 7. 8 and 7. 9 show the NSE performances for the different models assessed in both regions 1 and 2, respectively. The results are similar to those obtained by analyzing the RMSE performance. In most of the cases, ML obtained an NSE value above 0.9. The best values were obtained in JAE11 – Mancha Real (NSE = 0.963) in the first region and CAD04 - Villamartín (NSE = 0.926) in the second region, using MLP in both cases, whereas the worst model was carried out in JAE102 – Villacarrillo (NSE_{MLP} = 0.596, an NSE_{ELM} = 0.547 and NSE_{HS} = 0.459). On the other hand, there was only one site where HS outperformed ML models with a NSE difference higher than 0.05 ($NSE_{BestML-HS}$ = -0.054), GRA02 - Puebla de Don Fabrique. Moreover, the average NSE values carried out by ML models in both regions were very similar to one another (NSE_{Reg1} = 0.894 ± 0.064 and NSE_{Reg2} = 0.863 ± 0.060). Finally, considering the performance difference between HS and ML models, the mean outperformance in the second region was higher than in the second ($NSE_{BestML-HS}$ = 0.056 and $NSE_{BestML-HS}$ = 0.091 for regions 1 and 2, respectively).

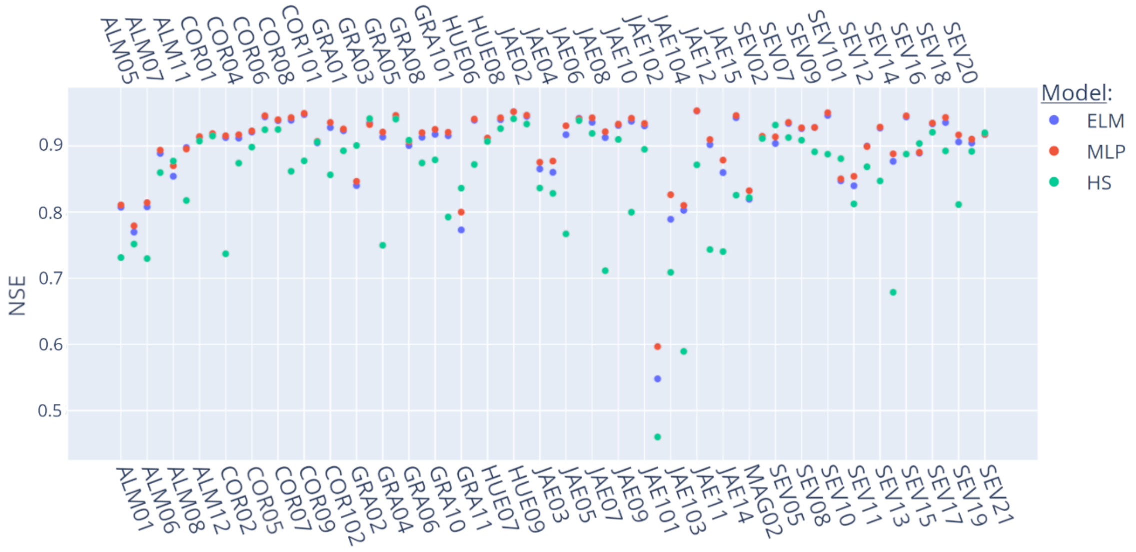


Figure 7. 8. NSE values for the best configurations in region 1.

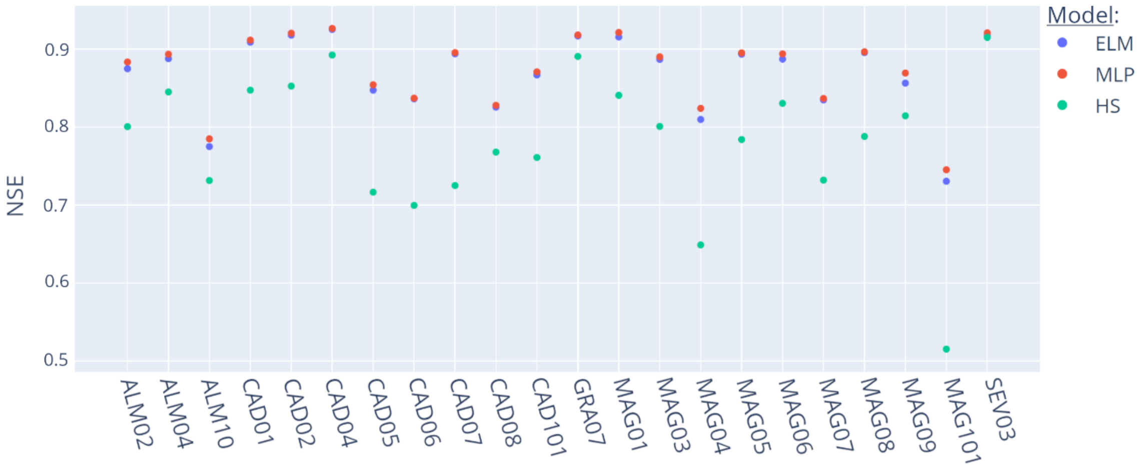


Figure 7. 9. NSE values for the best configurations in region 2.

Considering the R^2 values, Figures 10 and 11 show the R^2 values for the different models in the two assessed regions. The best ML value in the first region was performed in JAE05 - Torreblascopedro ($R^2 = 0.966$) using MLP whereas the worst was carried out in MAG02 – Vélez Málaga ($R^2 = 0.835$). In the second region, MAG05 - Archidona ($R^2 = 0.958$) got the best performance using MLP while ALM10 - Adra ($R^2 = 0.822$) was the worst. Overall, most of the AWSs performed above 0.9 in both regions. On average, the results carried out by ML models in the first region was higher than in the second region ($R^2 = 0.930 \pm 0.028$ and $R^2 = 0.895 \pm 0.038$, respectively).

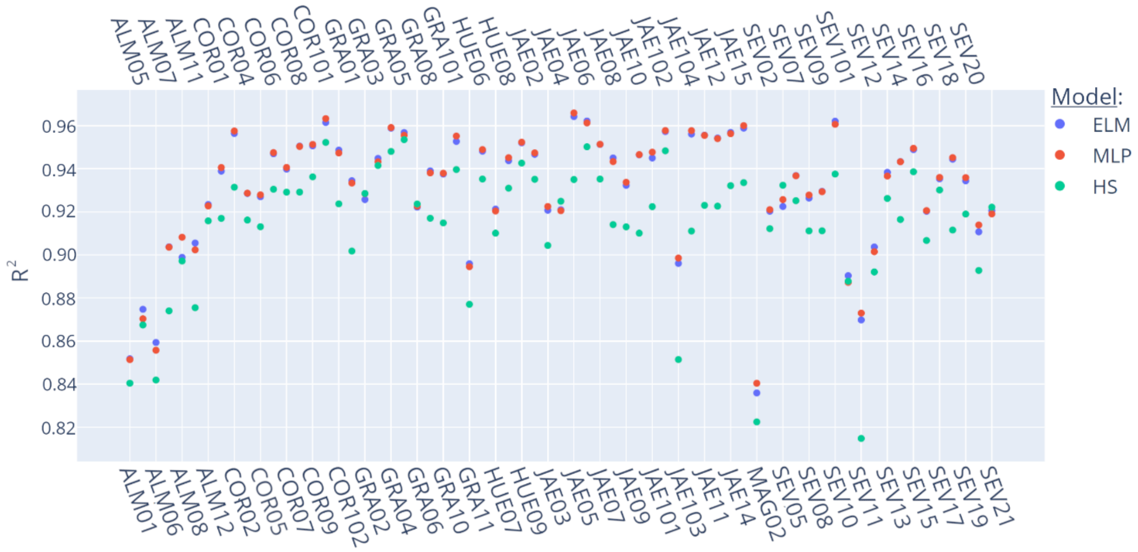


Figure 7. 10. R^2 values for the best configurations in region 1. See table 2 for further information about each location.

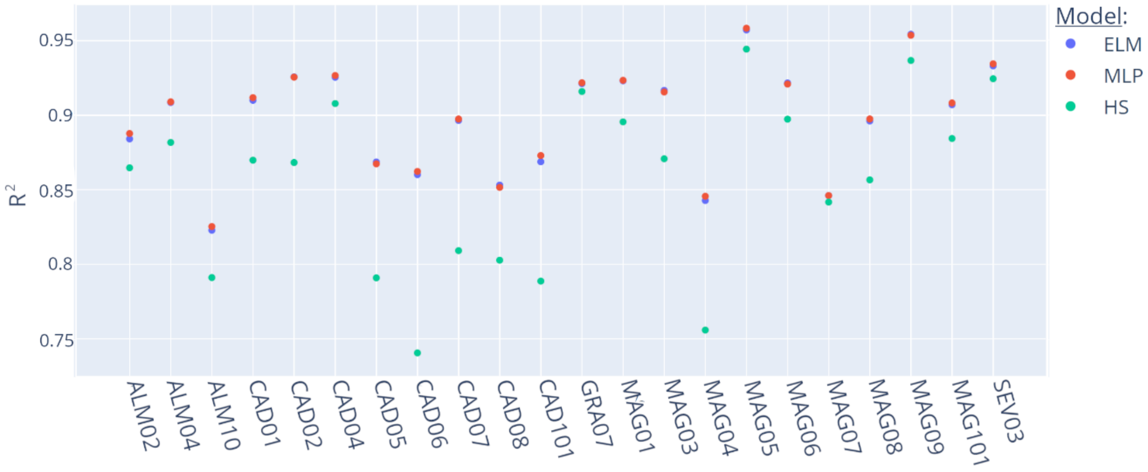


Figure 7. 11. R^2 values for the best configurations in region 2. See table 2 for further information about each location.

Finally, in terms of MBE, Figures 7. 12 and 7. 13 show the MBE values for the different models in the two analyzed regions. On average, the MBE performance in region 1 is MBE=0.099 ±0.363 mm/d whereas in the second region is MBE = 0.0327 ±0.307 mm/d. Besides, MLP generally outperformed the rest of the models in most of the cases, having the best value at 66 sites out of 89, whereas ELM outperformed HS and MLP only at 6 locations. The best value from the first region was carried out in SEV12 – La Rinconada (MBE=0.000 mm/d using ELM), JAE09 – Linares (MBE=0.002 mm/d using MLP) and GRA03 - Loja (MBE=0.011 mm/d using HS), while in the second region, CAD04 - Villamartín (MBE=-0.005 mm/d using ELM), ALM10 - Adra (MBE=0.0286 mm/d using MLP) and CAD01 - Basurta-Jerez de la Frontera (MBE=0.035 mm/d using MLP) performed as the best.

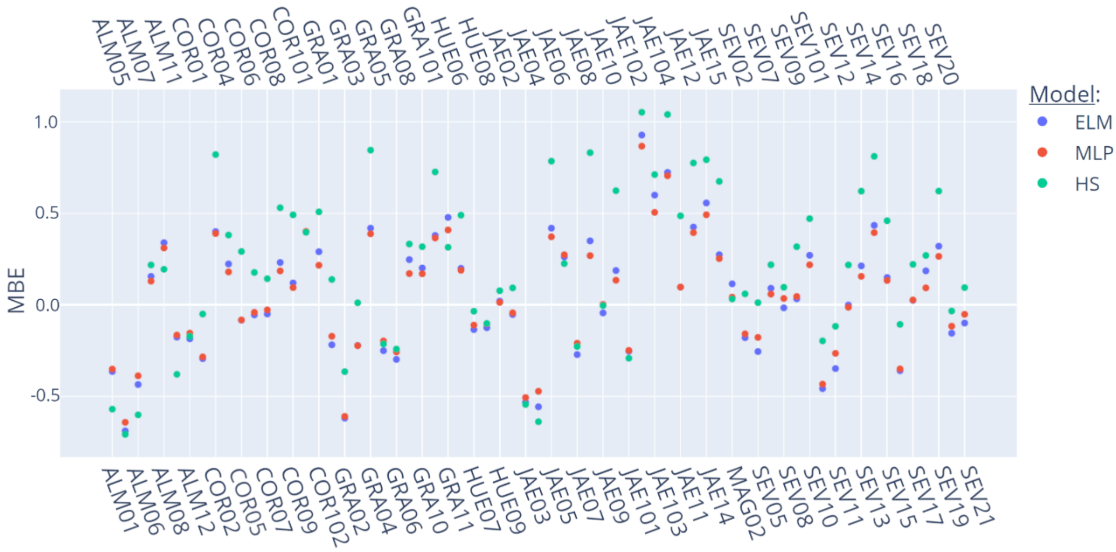


Figure 7. 12. MBE values for the best configurations in region 1. See table 2 for further information about each location.

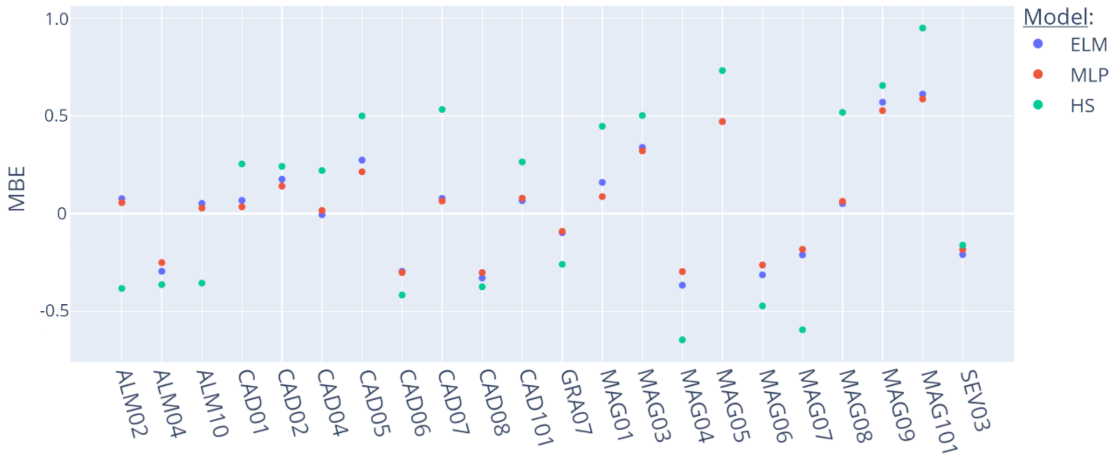


Figure 7. 13. MBE values for the best configurations in region 2. See table 2 for further information about each location.

All in all, the average RMSE, NSE, and R^2 values of ML models in the first region (RMSE=0.665 mm/d, NSE=0.894, and R^2= 0.930) were better than those obtained in the second (RMSE=0.712 mm/d, NSE=0.863, and R^2= 0.894), whereas, in terms of average MBE, both regions got a similar performance (MBE=0.040 mm/d). A possible justification for this behavior may be related to the effect of smoother DTR values in ET_0 modeling. As it has already been stated by other works in this same region [6,9], there is a lower accuracy in coastal sites modeling temperature-based ET_0 and Rs. Thus, modeling ET_0 in coastal zones (or inland sites with similar DTR multifractal characteristics) has a negative effect on the final model performance. Moreover, the improvements of ML models against HS were less significant in region 1 than those carried out in region 2, regarding mean RMSE ($RMSE_{BestML-HS}$ = -0.157 mm/d and $RMSE_{BestML-HS}$ = -0.212, respectively) and NSE ($NSE_{BestML-HS}$ = 0.056 and NSE =0.091, respectively). It seems to be the result of two main issues, the better performance of machine learning models compared to empirical equations and the higher information extraction from the temperature-based features $Energy_T$, HTx, and HTn, compared to only using Tx, Tn, and Ra (HS equation). Besides, MLP predominantly outperformed the rest of the models (ELM and HS) for the great majority of cases. However, the ML models underperformed in two sites from the first region (GRA02 – Puebla de Don Fadrique and SEV101 – IFAPA Centro de Las Torres-Tomejil-Finca Tomejil). One of the possible reasons may be the high wind speed values over the years (Figure 7. 14), with average values of 2.696 m/s and 2.408 m/s, respectively (Table 7. 2). As an example, wind speed data of two more stations have been included in this figure: JAE08 (Úbeda) and COR02 (Adamuz), with average values of 0.95 m/s and 1.03 m/s, respectively.

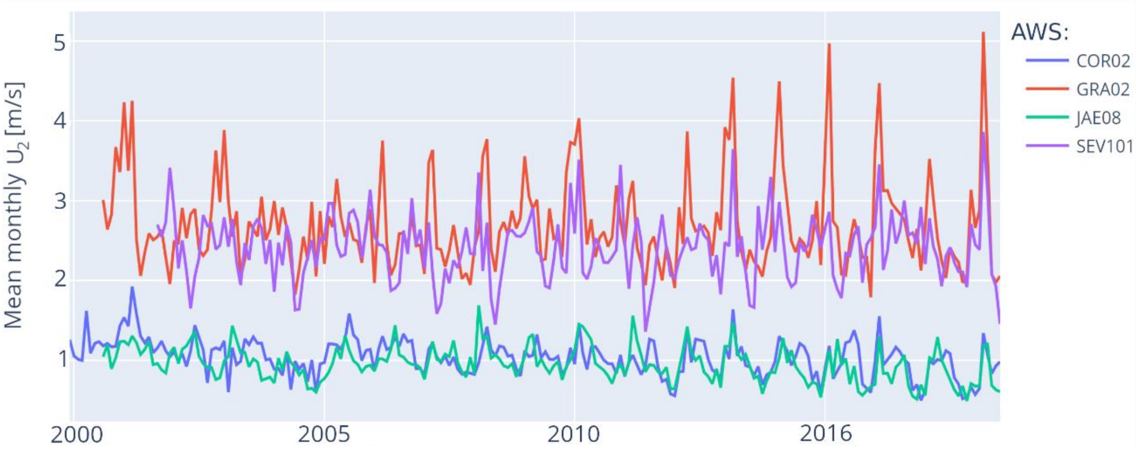


Figure 7. 14. Mean monthly wind speed values for GRA02 (Puebla de Don Fadrique), SEV101 (IFAPA Centro de Las Torres.-Tomejil-Finca Tomejil), JAE08 (Úbeda) and COR02 (Adamuz).

3.2.Assessment of configurations

In order to evaluate the different configurations, Table 7. 5 shows the minimum, mean and maximum values of RMSE, NSE, R^2 and MBE, as well as the GPI of the different configurations for all the AWS using MLP, due to its significant better results in the previous section. Generally, there were no big differences in performance between any of the ten configurations. In terms of RMSE, the configuration IV obtained the best performance ($RMSE_{min}$ = 0.4733 mm/d and $RMSE_{mean}$ = 0.6686 mm/d), followed very closely by configuration VI ($RMSE_{min}$ = 0.4756 mm/d and $RMSE_{mean}$ = 0.6731 mm/d). Regarding NSE, the configuration IV outperformed the rest of the configurations on average (NSE = 0.8892), while the configuration VIII carried out the highest value (NSE_{max} = 0.9534). Besides, in terms of R^2, configurations I and VI obtained the best mean performance (R^2 = 0.9224 and R^2 = 0.9224, respectively) whereas configuration 4 got the highest value (R^2 = 0.9659). In terms of MBE, configuration X obtained the minimum value (MBE = -0.0001 mm/d), while configuration IV carried out the best performance on average (MBE = 0.0377 mm/d) Finally, to evaluate all these statistics, a global performance indicator (GPI) was calculated, ranking configurations IV (GPI = 0.4290), I (GPI = 0.4236), V, VI and VII (GPI = 0.4190) as the best, in this order.

Table 7. 5. Statistical analysis regarding the different configurations for all the stations using MLP. The best values are in bold. RMSE represents Root Mean Square Error, NSE represents Nash-Sutcliffe Efficiency model, R^2 represents the Coefficient of Determination, MBE represents the Mean Bias Error and GPI represents the Global Performance Indicator.

		RMSE			**NSE**			**R^2**			**MBE**		
Conf	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**GPI**
I	0.4773	0.6721	1.1259	0.5507	0.8882	0.9532	0.8134	**0.9224**	0.9654	-0.0137	0.0479	0.9268	0.4236
II	0.4776	0.6761	1.0679	0.5966	0.8874	0.9515	0.8218	0.9216	0.9652	-0.0062	0.0427	0.8671	0.4168
III	0.4896	0.6790	1.2096	0.4814	0.8857	0.9533	0.8240	0.9216	0.9646	0.0082	0.0473	0.9900	0.4115
IV	**0.4733**	**0.6686**	1.1247	0.5516	**0.8892**	0.9524	0.8097	0.9222	**0.9659**	0.0073	**0.0377**	0.9234	**0.4290**
V	0.4791	0.6744	1.1902	0.4979	0.8869	0.9528	0.8200	0.9220	0.9641	0.0029	0.0456	0.9722	0.4190
VI	0.4756	0.6731	1.0735	0.5915	0.8882	0.9515	0.8131	**0.9224**	0.9651	-0.0009	0.0380	0.8784	0.4190
VII	0.4769	0.6784	1.1385	0.5405	0.8864	0.9528	0.8197	0.9215	0.9651	0.0074	0.0483	0.9272	0.4190
VIII	0.4892	0.6765	1.1257	0.5508	0.8873	**0.9534**	0.8245	0.9215	0.9651	-0.0061	0.0445	0.9135	0.4162
IX	0.4787	0.6819	1.2409	0.4542	0.8846	0.9527	0.8246	0.9217	0.9636	0.0030	0.0488	1.0101	0.4069
X	0.4788	0.6774	1.2566	0.4403	0.8858	0.9514	0.8253	0.9219	0.9641	**-0.0001**	0.0452	1.0666	0.4139

Table 7. 6 shows the same information as Table 7. 5, although they were calculated using only the results from the first region. Regarding RMSE, configuration IV outperformed the rest of the configurations in absolute and average terms ($RMSE_{min}$=0.4733 mm/d and $RMSE_{mean}$=0.6572 mm/d). In terms of NSE, configuration IV also carried out the best NSE performance on average (NSE=0.8967), whereas configuration VIII obtained the maximum value (NSE=0.9534). Regarding R^2, on average, configuration VI (R^2 = 0.9309) obtained the best performance, followed very closely by configurations IV and V (R^2=0.9306), whereas on maximum values, configurations IV (R^2 =0.9659), I (R^2=0.9654) and II (R^2=0.9652) were at the top of the ranking. In terms of MBE, all the minimum and mean performances were very accurate, obtaining values below 0.06 on average, being configuration IV the one that obtained the best average value (MBE=0.0379 mm/d). On the other hand, the best value was carried out by configuration VI (MBE=-0.0009 mm/d), closely followed by configuration VI (MBE=-0.0009 mm/d). Finally, regarding the GPI, configuration IV was at the top of the ranking (GPI=0.4330) and the configuration VI as the second (GPI=0.4250).

Table 7. 7 show the same information as Table 6, although they were calculated using the results from the second region. Regarding the RMSE performance, the configuration IV obtained the lowest minimum and mean values ($RMSE_{min}$ = 0.5481 mm/d and $RMSE_{mean}$ = 0.7034 mm/d). In terms of NSE, the maximum values were carried out by the configurations I (NSE = 0.9265) and IV (NSE = 0.9262), while on average configurations IV and X were the best (NSE = 0.8665). As regards the R^2 performance, on average, the top configurations were configuration I (R^2 = 0.8971), IV (R^2 = 0.8968) and VI (R^2 = 0.8964), whereas in absolute terms the configurations VI (R^2 = 0.9588), I (R^2 = 0.9584) and IV (R^2 = 0.9579) were the best. In terms of MBE, configurations VI (MBE = 0.0254 mm/d), IV (MBE = 0.0369 mm/d), VII (MBE = 0.0387 mm/d) and IX (MBE = 0.0357 mm/d) obtained MBE<0.4, on average. Besides, the best value was carried out by the configuration X (MBE = -0.0001 mm/d). Finally, regarding the GPI performance, the top configurations were configurations IV (GPI = 0.2945), I (GPI = 0.2885) and X (GPI = 0.2878), in this order.

To sum up, the configuration IV obtained the best GPI performance in both regions, highlighting accurate generalization of this configuration, as well as the importance of introducing $Energy_T$ (day), $Energy_T$ (day-1), HTx (day) and Tx (day-1) as a complement to the HS inputs to outperform ET_0 estimations.

Table 7. 6. Statistical analysis regarding the different configurations for AWSs from region 1 using MLP. RMSE represents Root Mean Square Error, NSE represents Nash-Sutcliffe Efficiency model, R^2 represents the Coefficient of Determination, MBE represents the Mean Bias Error and GPI represents the Global Performance Indicator.

		RMSE			NSE			R^2			MBE		
Conf	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**GPI**
I	0.4773	0.6611	1.1259	0.5507	0.8954	0.9532	0.8376	0.9308	0.9654	-0.0137	0.0497	0.9268	0.4139
II	0.4776	0.6662	1.0679	0.5966	0.8944	0.9515	0.8339	0.9300	0.9652	-0.0062	0.0423	0.8671	0.4186
III	0.4896	0.6683	1.2096	0.4814	0.8927	0.9533	0.8367	0.9300	0.9646	0.0101	0.0481	0.9900	0.4142
IV	**0.4733**	**0.6572**	1.1247	0.5516	**0.8967**	0.9524	0.8351	0.9306	**0.9659**	0.0073	**0.0379**	0.9234	**0.4330**
V	0.4791	0.6618	1.1902	0.4979	0.8946	0.9528	0.8417	0.9306	0.9641	0.0029	0.0453	0.9722	0.4250
VI	0.4756	0.6611	1.0735	0.5915	0.8958	0.9515	0.8394	**0.9309**	0.9651	**-0.0009**	0.0421	0.8784	0.4274
VII	0.4769	0.6672	1.1385	0.5405	0.8935	0.9528	0.8360	0.9299	0.9651	0.0074	0.0515	0.9272	0.4163
VIII	0.4892	0.6652	1.1257	0.5508	0.8947	**0.9534**	0.8338	0.9299	0.9651	-0.0086	0.0450	0.9135	0.4201
IX	0.4787	0.6730	1.2409	0.4542	0.8910	0.9527	0.8361	0.9302	0.9636	0.0030	0.0532	1.0101	0.4067
X	0.4788	0.6681	1.2566	0.4403	0.8921	0.9514	0.8359	0.9302	0.9641	-0.0070	0.0452	1.0666	0.4141

Table 7. 7. Statistical analysis regarding the different configurations for AWSs from region 2 using MLP. RMSE represents Root Mean Square Error, NSE represents Nash-Sutcliffe Efficiency model, R^2 represents the Coefficient of Determination, MBE represents the Mean Bias Error and GPI represents the Global Performance Indicator

	RMSE			**NSE**			**R^2**			**MBE**			
Conf	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**GPI**
I	0.5635	0.7057	0.9148	0.7288	0.8662	**0.9265**	0.8134	**0.8971**	0.9584	0.0164	0.0425	0.6218	0.2885
II	0.5560	0.7065	0.9157	0.7235	0.8661	0.9243	0.8218	0.8960	0.9570	-0.0166	0.0436	0.6381	0.2850
III	0.5692	0.7116	0.8900	0.7191	0.8644	0.9239	0.8240	0.8961	0.8961	0.0082	0.0449	0.6303	0.2699
IV	**0.5481**	**0.7034**	0.8986	0.7104	**0.8665**	0.9262	0.8097	0.8968	0.9579	-0.0220	0.0369	0.6531	**0.2945**
V	0.5663	0.7128	0.9030	0.7112	0.8635	0.9251	0.8200	0.8958	0.9561	-0.0128	0.0466	0.6536	0.2655
VI	0.5566	0.7098	0.8841	0.7275	0.8651	0.9241	0.8131	0.8964	**0.9588**	0.0087	**0.0254**	0.6158	0.2756
VII	0.5723	0.7123	0.8918	0.7453	0.8646	0.9208	0.8197	0.8959	0.9575	0.0286	0.0387	0.5862	0.2679
VIII	0.5591	0.7110	0.8968	0.7197	0.8645	0.9227	0.8245	0.8959	0.9573	-0.0061	0.0429	0.6276	0.2714
IX	0.5592	0.7091	0.9214	0.7136	0.8652	0.9234	0.8246	0.8959	0.9574	0.0224	0.0357	0.6369	0.2771
X	0.5567	0.7059	0.9087	0.7141	**0.8665**	0.9240	0.8253	0.8966	0.9573	**-0.0001**	0.0451	0.6449	0.2878

3.3.Seasonal performance

Due to the good results of configuration IV for both regions, the following analysis has been carried out only using this configuration as well as the MLP model. The following figures show the RMSE (Figure 7. 15), NSE (Figure 7. 16), R^2 (Figure 7. 17) and MBE (Figure 7. 18) for winter, summer, spring and autumn in both regions. Generally, the results obtained in region 1 outperformed those obtained in region 2. Besides, the estimations carried out in winter and autumn were better than those calculated in summer and spring.

Moreover, specifically, regarding RMSE (Figure 7. 17), winter obtained the best performance (RMSE = 0.307 mm/d) in JAE06 (Alcaudete), whereas the worst value was obtained in summer (RMSE = 1.414 mm/d) in SEV11 (Osuna), both carried out in the first region. On average, the results of AWSs from the first region outperformed those obtained in the second region, having a mean difference of 0.075 mm/d between the different regions and seasons. In terms of NSE, the best results were carried out in autumn (NSE=0.878) in JAE101 (Torreperogil) whereas the worst value was also obtained in summer (NSE=-0.77) in MAG101 (IFAPA Centro de Campanillas). On average, spring and autumn obtained the most accurate performances, followed by winter and summer, in this order. Regarding the R^2, autumn and winter carried out the best value in absolute terms (R^2 = 0.937 and R^2 = 0.922, respectively), whereas on average, autumn and spring were on top. Besides, as it is common in the aforementioned statistics (RMSE, NSE and R^2), the results given by the first region outperformed those obtained in the second region. Eventually, regarding the MBE, the results obtained by both regions were very accurate on average (MBE values close to 0) with no significant differences from one another. However, the interquartile range of autumn and winter was lower than in spring and summer, denoting a more precise estimation of the first two seasons.

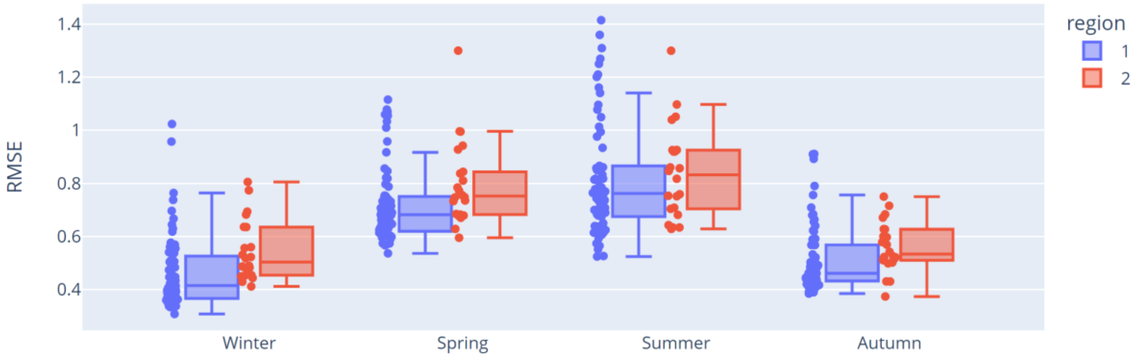


Figure 7. 15. RMSE (Root Mean Square Error) seasonal values for all the AWS using configuration 4.

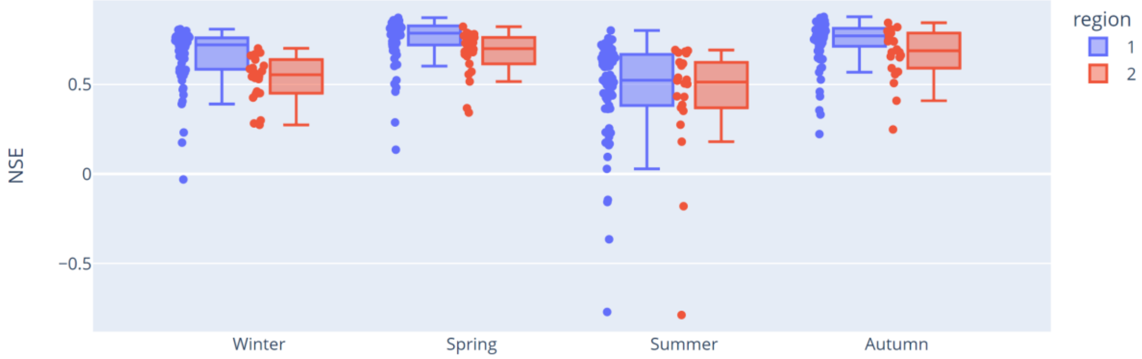


Figure 7. 16. . NSE (Nash-Sutcliffe Efficiency model) seasonal values for all the AWS using configuration 4.

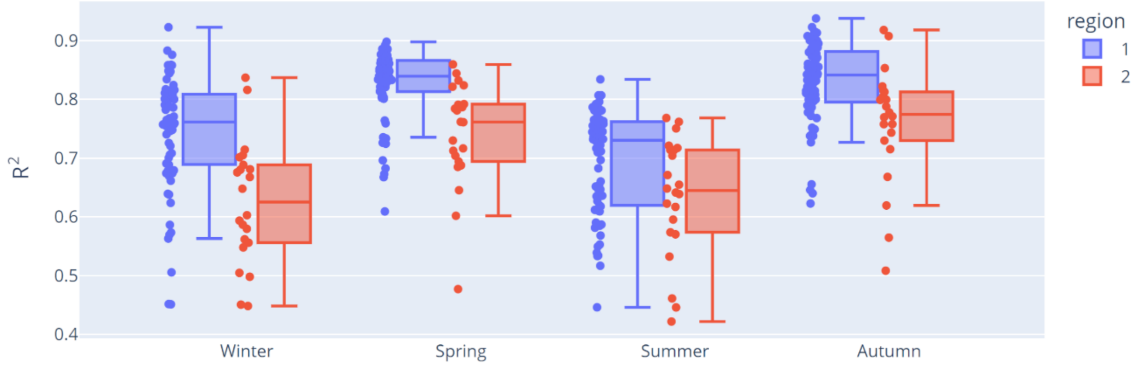


Figure 7. 17. R^2 (Coefficient of Determination) seasonal values for all the AWS using configuration 4.

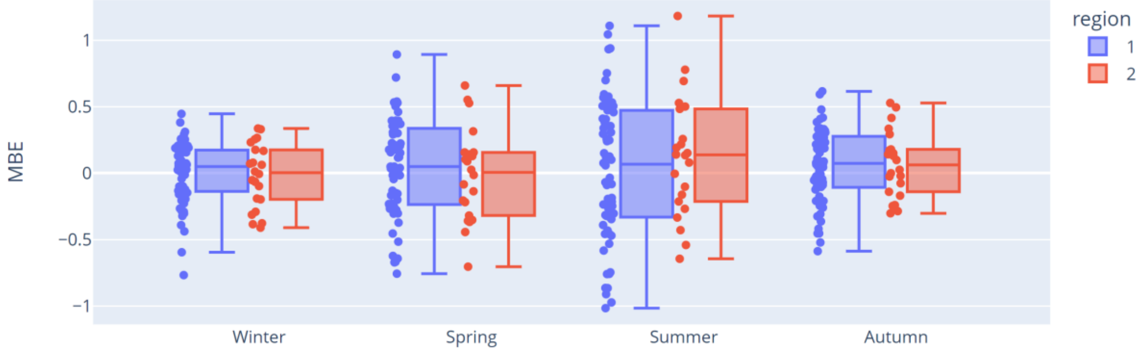


Figure 7. 18. MBE (Mean Bias Error) seasonal values for all the AWS using configuration 4.

3.4 Overall discussion

In the first place, comparing the regional approach developed in this work with the local-calibration carried out in this same region by Bellido-Jiménez et al. (2020), the results were promising. For example, in MAG01, the regional models outperformed local-calibration in terms of RMSE, NSE and R^2 ($RMSE_{regional}$ = 0.5481 mm/d, $RMSE_{local}$ = 0.5952 mm/d, $NSE_{regional}$ = 0.9216, NSE_{local} = 0.9091, $R^2_{regional}$ = 0.9239 and R^2_{local} = 0.9105), whereas in the rest of the sites, the results were very close one another in terms of accuracy. Besides, the results obtained in this work outperformed those carried out by Ferreira et al. [22], where ET_0 was estimated using sub-daily temperature (24 hourly values as features) and the region was randomly generated across the whole study area, the state of Minas Gerais (Brazil). In terms of NSE an R^2, on average, MLP using the configuration 4 outperformed the results of CNN in Brazil (NSE_{Brazil} = 0.83, R^2_{Brazil} = 0.87, $NSE_{Region\ 1}$ = 0.89, $R^2_{Region\ 1}$ = 0.93, $NSE_{Region\ 2}$ = 0.87 and $R^2_{Region\ 2}$ = 0.89). However, the mean RMSE performance was slightly better in Brazil, obtaining an RMSE = 0.52 mm/d, whereas in this work the mean values were $RMSE_{Region\ 1}$ = 0.65 mm/d and $RMSE_{Region\ 2}$ = 0.70 mm/d. Moreover, the results of this study also outperformed those obtained by Kazemi et al [61], where the training dataset was composed of 10 stations from Turkey using GEP and the models were assessed in 18 locations in Iran with different climate aridity conditions (6 arid locations, 6 semi-arid locations, and 6 humid locations). On average, the results from the arid, semi-arid, and humid locations were R^2=0.823, R^2=0.888, and R^2=0.859, respectively. Finally, the results from this work were very similar to those carried out by Yan et al. [62] where meteorological data from adjacent stations were used for training. The results in terms of RMSE and NSE were, on average, RMSE =0.685 mm/d and NSE = 0.922 for the arid locations and RMSE = 0.6425 mm/d and NSE = 0.849 for humid sites.

In general, comparing the results from the two different regions, those carried out in sites from the first region were better in all statistics than those obtained in AWSs from the second, where the first region got an RMSE = 0.6572 mm/d, NSE = 0.8967, R^2 = 0.9306 and MBE = |0.05|mm/d and the second, RMSE = 0.72 mm/d, NSE = 0.85, R^2 = 0.89 and MBE = |0.04|mm/d. A possible justification may be related to the effect of smoother DTR values for ET_0 modeling, the same as it happens in coastal sites for this same region [6,9]. Moreover, seasonally, regarding RMSE, NSE, and R^2 the models deployed in region 1 also outperformed those carried out in the second in all the seasons, being spring and autumn the ones that obtained the best results in terms of NSE and R^2. However, in terms of RMSE, winter outperformed the rest of the seasons, although it is related to the lower ET_0 values during this period in this region.

Besides, ML highly outperformed HS in most of the cases, especially MLP, which could be the result of two main issues, the better performance of machine learning models compared to empirical equations and the higher information extraction from the temperature-based features $Energy_T$, HTx, and HTn, compared to only using Tx, Tn, and Ra (HS equation). It is worth noting that configuration IV, which is composed of the variables Tx (day -1), Tx (day), Tn (day), Ra (day), $Energy_T$ (day-1), $Energy_T$ (day), and HTx (day), obtained the best performance in both regions, carrying out a GPI performance of GPI=0.4330 in the first region and a GPI=0.2945 in the second region. It states the valuable information of EnergyT and HTx for temperature-based ET_0 estimations.

Furthermore, the improvements of MLP against HS were more significant in the second region in terms of RMSE ($RMSE_{BestML-HS}$ = -0.157 mm/d and $RMSE_{BestML-HS}$ = -0.212 mm/d for the first and second region respectively), NSE, and R^2 (up to a 37.5% for the first region and up to 44.7% for the second region), whereas regarding MBE, there was no such significant difference between them. Nevertheless, there were 2 sites (GRA02 and SEV101) where ML underperformed HS in terms of RMSE and NSE by around 0.1 mm/d, whereas regarding R^2, ML models were always superior. This weird low accuracy in ML models seems to be related to high wind speed values. Moreover, regarding the Bayesian optimization, deeper MLP architectures (more than 3 hidden layers and more than 15 neurons per layer) were selected as the fittest for this task due to the higher number of inputs compared to other literature works addressing temperature-based approaches (only using 3 or 4 inputs such as Ra, Tx, Tn, and Tm), which is the cause of the higher computational cost of MLP against ELM (more than 3 million weights to update during the backpropagation process in the worse scenario). A possible limitation in this work is the strong dependence on accurate and representative temperature records since the models exclusively depend on temperature values. Besides, despite Andalusia being a very geoclimatic heterogeneous region, it would be interesting to assess these two models in different regions with different characteristics such as extreme aridity index locations. Finally, the use of unsupervised machine learning approaches such as K-means cluster can be further evaluated, as well as the use of more variables such as RH, u_2, and Rs.

4. Conclusions

In this work, the main objective was to come up with a novel approach to temperature-based ET_0 models to deal with the problem of requiring in situ large datasets to build accurate ML models. In this sense, a new regionalization has been carried out based on multifractal characteristics of DTR values into two main regions. Thus, it allows a more accurate station grouping regardless of their geographical distribution, only using temperature values for this purpose.

Up to 70 different configurations have been assessed using combinations of Tx, Tn, $Energy_T$, HTx, and HTn from the same and previous day as the target, but only the best 10 input configurations were evaluated in this work. The configuration composed of the variables Tx (day -1), Tx (day), Tn (day), Ra (day), $Energy_T$ (day-1), $Energy_T$ (day), and HTx (day) was the one that obtained the best modeling in both regions. Moreover, MLP highly outperformed the rest of the models in terms of RMSE, NSE, and R^2, whereas regarding MBE, no such significant differences were appreciated. The results of MLP

Due to the results of this work it could be stated that ET_0 models deployed in areas from the first region were more accurate than those carried out in the second. However, relatively, the outperformance of MLP against HS was more significant than AWS sites from the second region. Finally, the approaches proposed in this work may result in greater efficiency for optimizing water resources, allowing to accurately estimate ET_0 only using temperature records.

5. References

[1] Peña-Gallardo M, Gámiz-Fortis SR, Castro-Díez Y, Esteban-Parra MJ. Análisis comparativo de índices de sequía en andalucía para el periodo 1901-2012. Cuadernos de Investigacion Geografica 2016;42:67–88. https://doi.org/10.18172/cig.2946.

[2] Estévez J, Bellido-Jiménez JA, Liu X, García-Marín AP. Monthly Precipitation Forecasts Using Wavelet Neural Networks Models in a Semiarid Environment. Water 2020;12:1909. https://doi.org/10.3390/w12071909.

[3] Allen R, Pereira L, Smith M. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage. vol. 56. 1998.

[4] Estévez J, P.Gavilán, J.V.Giráldez. Guidelines on validation procedures for meteorological data from automatic weather stations. J Hydrol (Amst) 2011;402:144–54. https://doi.org/10.1016/j.jhydrol.2011.02.031.

[5] Estévez J, García-Marín AP, Morábito JA, Cavagnaro M. Quality assurance procedures for validating meteorological input variables of reference evapotranspiration in mendoza province (Argentina). Agricultural Water Management 2016;172:96–109. https://doi.org/10.1016/j.agwat.2016.04.019.

[6] Estévez J, Padilla FL, Gavilán P. Evaluation and Regional Calibration of Solar Radiation Prediction Models in Southern Spain. Journal of Irrigation and Drainage Engineering 2012;138:868–79. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000493.

[7] WMO WMO. Guide to Instruments and Methods of Observations. vol. 8. Geneva, Switzerland.: WMO; 2018.

[8] Hargreaves GH, Samani ZA. Reference Crop Evapotranspiration from Temperature. Applied Engineering in Agriculture 1985;1:96–9. https://doi.org/10.13031/2013.26773.

[9] Bellido-Jiménez JA, Estévez J, García-Marín AP. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agric Water Manag 2020;245:106558. https://doi.org/10.1016/j.agwat.2020.106558.

[10] Zhou Z, Zhao L, Lin A, Qin W, Lu Y, Li J, et al. Exploring the potential of deep factorization machine and various gradient boosting models in modeling daily reference evapotranspiration in China. Arabian Journal of Geosciences 2020;13:1–20. https://doi.org/10.1007/s12517-020-06293-8.

[11] Muhammad MKI, Shahid S, Ismail T, Harun S, Kisi O, Yaseen ZM. The development of evolutionary computing model for simulating reference evapotranspiration over Peninsular Malaysia. Theor Appl Climatol 2021;144:1419–34. https://doi.org/10.1007/s00704-021-03606-z.

[12] Adnan RM, Mostafa R, Islam ARMT, Kisi O, Kuriqi A, Heddam S. Estimating reference evapotranspiration using hybrid adaptive fuzzy inferencing coupled with heuristic algorithms. Comput Electron Agric 2021;191:106541. https://doi.org/10.1016/j.compag.2021.106541.

[13] Sattari MT, Apaydin H, Band SS, Mosavi A, Prasad R. Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrol Earth Syst Sci 2021;25:603–18. https://doi.org/10.5194/hess-25-603-2021.

[14] Bedi J. Transfer learning augmented enhanced memory network models for reference evapotranspiration estimation. Knowl Based Syst 2022;237:107717. https://doi.org/10.1016/j.knosys.2021.107717.

[15] Maroufpoor S, Bozorg-Haddad O, Maroufpoor E. Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. J Hydrol (Amst) 2020;588:125060. https://doi.org/10.1016/j.jhydrol.2020.125060.

[16] Malik A, Kumar A, Kim S, Kashani MH, Karimi V, Sharafati A, et al. Modeling monthly pan evaporation process over the Indian central Himalayas: application of multiple learning artificial intelligence model. Http://WwwTandfonlineCom/Action/AuthorSubmission?JournalCode=tcfm20&page=instructions 2020;14:323–38. https://doi.org/10.1080/19942060.2020.1715845.

[17] Dimitriadou S, Nikolakopoulos KG. Annual Actual Evapotranspiration Estimation via GIS Models of Three Empirical Methods Employing Remotely Sensed Data for the Peloponnese, Greece, and Comparison with Annual MODIS ET and Pan Evaporation Measurements. ISPRS International Journal of Geo-Information 2021, Vol 10, Page 522 2021;10:522. https://doi.org/10.3390/IJGI10080522.

[18] Gavilán P, Lorite IJ, Tornero S, Berengena J. Regional calibration of Hargreaves equation for estimating reference ET in a semiarid environment. Agric Water Manag 2006;81:257–81. https://doi.org/https://doi.org/10.1016/j.agwat.2005.05.001.

[19] Dai X, Shi H, Li Y, Ouyang Z, Huo Z. Artificial neural network models for estimating regional reference evapotranspiration based on climate factors. Hydrol Process 2009;23:442–50. https://doi.org/10.1002/hyp.7153.

[20] Ramírez-Cuesta JM, Cruz-Blanco M, Santos C, Lorite IJ. Assessing reference evapotranspiration at regional scale based on remote sensing, weather forecast and GIS tools. International Journal of Applied Earth Observation and Geoinformation 2017;55:32–42. https://doi.org/10.1016/j.jag.2016.10.004.

[21] Feng Y, Cui N, Gong D, Zhang Q, Zhao L. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. Agric Water Manag 2017;193:163–73. https://doi.org/10.1016/j.agwat.2017.08.003.

[22] Ferreira LB, da Cunha FF. New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 2020;234:106113. https://doi.org/https://doi.org/10.1016/j.agwat.2020.106113.

[23] Trajkovic S, Gocic M, Pongracz R, Bartholy J. Adjustment of Thornthwaite equation for estimating evapotranspiration in Vojvodina. Theor Appl Climatol 2019;138:1231–40. https://doi.org/10.1007/s00704-019-02873-1.

[24] Fang W, Huang S, Huang Q, Huang G, Meng E, Luan J. Reference evapotranspiration forecasting based on local meteorological and global climate information screened by partial mutual information. J Hydrol (Amst) 2018;561:764–79. https://doi.org/10.1016/j.jhydrol.2018.04.038.

[25] Wu L, Peng Y, Fan J, Wang Y, Huang G. A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating monthly reference evapotranspiration in parallel computation. Agric Water Manag 2021;245:106624. https://doi.org/10.1016/j.agwat.2020.106624.

[26] Ferreira LB, da Cunha FF. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. Comput Electron Agric 2020;234:106113. https://doi.org/https://doi.org/10.1016/j.agwat.2020.106113.

[27] Shiri J, Nazemi AH, Sadraddini AA, Landeras G, Kisi O, Fakheri Fard A, et al. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Computers and Electronics in Agriculture 2014;108:230–41. https://doi.org/10.1016/j.compag.2014.08.007.

[28] Paredes P, Pereira LS. Computing FAO56 reference grass evapotranspiration PM-ET o from temperature with focus on solar radiation. Agricultural Water Management 2019;215:86–102. https://doi.org/10.1016/j.agwat.2018.12.014.

[29] Herrera-Grimaldi P, García-Marín AP, Estévez J. Multifractal analysis of diurnal temperature range over Southern Spain using validated datasets. Chaos 2019;29:063105. https://doi.org/10.1063/1.5089810.

[30] Bellido-Jiménez JAntonio, Estévez J, Vanschoren J, García-Marín APenélope; AgroML: An Open-Source Repository to Forecast Reference Evapotranspiration in Different Geo-Climatic Conditions Using Machine Learning and Transformer-Based Models. Agronomy 2022;12:656. https://doi.org/10.3390/agronomy12030656.

[31] Bellido-Jiménez JA, Estévez Gualda J, García-Marín AP. Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions. Applied Energy 2021;298:117211. https://doi.org/10.1016/j.apenergy.2021.117211.

[32] Schertzer D, Lovejoy S. Physical modeling and analysis of rain and clouds by anisotropic scaling mutiplicative processes. Journal of Geophysical Research 1987;92:9693–714. https://doi.org/10.1029/JD092iD08p09693.

[33] De Lima MIP, De Lima JLMP. Investigating the multifractality of point precipitation in the Madeira archipelago. Nonlinear Process Geophys 2009;16:299–311. https://doi.org/10.5194/npg-16-299-2009.

[34] García-Marín AP, Ayuso-Muñoz JL, Jiménez-Hornero FJ, Estévez J. Selecting the best IDF model by using the multifractal approach. Hydrolog Process 2013;27:433–43. https://doi.org/10.1002/hyp.9272.

[35] Herrera-Grimaldi P, García-Marín A, Ayuso-Muñoz JL, Flamini A, Morbidelli R, Ayuso-Ruíz JL. Detection of trends and break points in temperature: the case of Umbria (Italy) and Guadalquivir Valley (Spain). Acta Geophysica 2018;66:329–43. https://doi.org/10.1007/s11600-018-0118-1.

[36] Estévez J, Gavilán P, García-Marín AP. Spatial regression test for ensuring temperature data quality in southern Spain. Theor Appl Climatol 2018;131:309–18. https://doi.org/10.1007/S00704-016-1982-8.

[37] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric For Meteorol 1984;31:159–66. https://doi.org/10.1016/0168-1923(84)90017-0.

[38] Hossein Kazemi M, Shiri J, Marti P, Majnooni-Heris A. Assessing temporal data partitioning scenarios for estimating reference evapotranspiration with machine learning techniques in arid regions. Journal of Hydrology 2020;590:125252. https://doi.org/https://doi.org/10.1016/j.jhydrol.2020.125252.

[39] Shiri J, Marti P, Karimi S, Landeras G. Data splitting strategies for improving data driven models for reference evapotranspiration estimation among similar stations. Computers and Electronics in Agriculture 2019;162:70–81. https://doi.org/https://doi.org/10.1016/j.compag.2019.03.030.

[40] Paola F, Giugni M. Coupled Spatial Distribution of Rainfall and Temperature in USA. Procedia Environmental Sciences 2013;19:178–187. https://doi.org/https://doi.org/10.1016/j.proenv.2013.06.020.

[41] Yin J, Deng Z, Ines AVM, Wu J, Rasu E. Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM). Agricultural Water Management 2020;242:106386. https://doi.org/10.1016/J.AGWAT.2020.106386.

[42] Luo Y, Chang X, Peng S, Khan S, Wang W, Zheng Q, et al. Short-term forecasting of daily reference evapotranspiration using the Hargreaves-Samani model and temperature forecasts. Agricultural Water Management 2014;136:42–51. https://doi.org/10.1016/j.agwat.2014.01.006.

[43] Üneş F, Kaya YZ, Mamak M. Daily reference evapotranspiration prediction based on climatic conditions applying different data mining techniques and empirical equations. Theoretical and Applied Climatology 2020;141:763–73. https://doi.org/10.1007/s00704-020-03225-0.

[44] Thornthwaite CW, Mather JR. Instructions and tables for computing potential evapotranspiration and the water balance. Laboratory of Climatology; 1957.

[45] Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. Journal of Hydrology 2019;572:556–70. https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.03.028.

[46] Antonopoulos VZ, Antonopoulos A V. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. Computers and Electronics in Agriculture 2017;132:86–96. https://doi.org/10.1016/J.COMPAG.2016.11.011.

[47] Yassin MA, Alazba AA, Mattar MA. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. Agricultural Water Management 2016;163:110–24. https://doi.org/10.1016/j.agwat.2015.09.009.

[48] Guang-Bin H, Qin-Yu Z, Chee-Kheong S. Extreme learning machine: a new learning scheme of feedforward neural networks. 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), vol. 2, 2004, p. 985–90 vol.2. https://doi.org/10.1109/IJCNN.2004.1380068.

[49] Kisi O, Alizamir M. Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: Wavelet extreme learning machine vs wavelet neural networks. Agricultural and Forest Meteorology 2018;263:41–8. https://doi.org/10.1016/j.agrformet.2018.08.007.

[50] Borji A, Itti L. Bayesian optimization explains human active search. Advances in Neural Information Processing Systems 2013.

[51] Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.

[52] Tikhamarine Y, Malik A, Souag-Gamane D, Kisi O. Artificial intelligence models versus empirical equations for modeling monthly reference evapotranspiration. Environmental Science and Pollution Research 2020;27:30001–19. https://doi.org/10.1007/s11356-020-08792-3.

[53] Alizamir M, Kisi O, Muhammad Adnan R, Kuriqi A. Modelling reference evapotranspiration by combining neuro-fuzzy and evolutionary strategies. Acta Geophysica 2020;68:1113–26. https://doi.org/10.1007/s11600-020-00446-9.

[54] Roy DK, Barzegar R, Quilty J, Adamowski J. Using ensembles of adaptive neuro-fuzzy inference system and optimization algorithms to predict reference evapotranspiration in subtropical climatic zones. Journal of Hydrology 2020;591:125509. https://doi.org/10.1016/j.jhydrol.2020.125509.

[55] Kotthoff L, Thornton C, Hoos H, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research 2017;18:1–5.

[56] Jin H, Song Q, Hu X. Auto-Keras: An Efficient Neural Architecture Search System. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019:1946–1956. https://doi.org/10.1145/3292500.3330648.

[57] Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Auto-sklearn:: Efficient and robust automated machine learning. Advances in Neural Information Processing Systems, vol. 2015- Janua, 2015, p. 2962–70.

[58] Hutter F, Kotthoff L, Vanschoren J. Automated Machine Learning. In: Hutter F, Kotthoff L, Vanschoren J, editors., Cham: Springer International Publishing; 2019, p. 245–81. https://doi.org/10.1007/978-981-16-2233-5_11.

[59] Bateni SM, Khoshkam H, Valipour M. Forecasting daily reference evapotranspiration using hybrid Gaussian Programming Regression-Bayesian Optimization (GPR-BO) and Wavelet-GPR-BO. AGUFM 2021;2021:H21B-04.

[60] Alibabaei K, Gaspar PD, Lima TM. Modeling Soil Water Content and Reference Evapotranspiration from Climate Data Using Deep Learning Method. Applied Sciences 2021, Vol 11, Page 5029 2021;11:5029. https://doi.org/10.3390/APP11115029.

[61] Kazemi MH, Majnooni-Heris A, Kisi O, Shiri J. Generalized gene expression programming models for estimating reference evapotranspiration through cross-station assessment and exogenous data supply. Environmental Science and Pollution Research 2021;28:6520–32. https://doi.org/10.1007/s11356-020-10916-8.

[62] Yan S, Wu L, Fan J, Zhang F, Zou Y, Wu Y. A novel hybrid WOA-XGB model for estimating daily reference evapotranspiration using local and external meteorological data: Applications in arid and humid regions of China. Agric Water Manag 2021;244:106594. https://doi.org/10.1016/j.agwat.2020.106594.

Chapter 8

Conclusions

The Ph.D. has developed several strategies to outperform agrometeorological estimations and predictions, focusing not only on enhancing accuracy but also on minimizing the number of required parameters. Thus, only temperature-based approaches have been assessed, as temperature is the most widely used metric in weather stations, as well as being the cheapest and most reliable measure, even when measured with low-cost devices. Due to their importance in sectors such as agronomy, hydrology or energy sector, among others, the modeled agrometeorological variables have been precipitation, solar radiation, and reference evapotranspiration.

Firstly, novel temperature-based features have been evaluated such as DTR, $Energy_T$, HTx, HTn, HSr, and HSs, among others, calculated from hourly or intra-hourly temperature records. They have been assessed to outperform reference evapotranspiration and solar radiation predictions on both local and regional calibration approaches (papers 2, 3, 5, and 6). These new parameters have been proven to highly outperform empirical methods and all baselines in most cases.

Secondly, different machine learning models have been tested on different scenarios, such as MLP, ELM, SVM, RF, and XGBoost, among others. There exist differences in performance depending on the nature of the target variable (precipitation, solar radiation, and reference evapotranspiration) and other factors like the aridity index or the proximity to the sea. Roughly, it could be stated that MLP seems to obtain the most accurate results for precipitation estimations (papers 1 and 4), SVM generally outperforms MLP when estimating solar radiation (papers 3), and MLP and ELM are the fittest when predicting and forecasting ET_0 (papers 2, 5 and 6). It is worth noting that models such as RF, XGBoost, and Transformed-based models, despite having a very high computational cost, do not improve the rest of the models, getting even worse performance in some cases.

Thirdly, different approaches in terms of training have been evaluated. On the one hand, a local calibration algorithm obtains the best results generally (papers 1, 2, 3, and 4) since the models are trained on the same climatic conditions and site. This may give rise to the problem of lack of generalization in models, as these models are designed to be used in a specific location. However, deploying them in new and unseen locations with similar geoclimatic characteristics can still yield accurate estimations (paper 3). On the other hand, a regional approach was evaluated in the region of Andalusia, clustering the AWS according to the multifractals characteristics of DTR values (two main regions, inland sites - region 1 and coastal sites – region 2). The results highlight that the utilization of a regional approach highly enhances the performance of empirical methodologies and, despite being usually less efficient, it can outperform local calibration models in some cases.

Finally, as a significant milestone, an open-source repository (AgroML) has been developed. It provides researchers with a powerful tool to develop and compare various machine learning algorithms, including Random Forest, Extreme Learning Machine, Support Vector Machine, Multilayer Perceptron, Convolutional Neural Networks, Long Short-Term Memory, and Transformer-based models, for different regression applications. The repository has automated hyperparameters tuning using Bayesian optimization, ensuring a fair comparison between models and leading to more accurate and reliable results. Moreover, AgroML's functionality extends beyond standard machine learning applications, with the inclusion of specific functions such as the ET_0 empirical formula and splitting functions for AWS. These features enhance the repository's ability to tackle complex agricultural problems, making it a valuable resource for technicians, farmers, researchers and practitioners alike. Overall, AgroML's open-source nature, the wide range of machine learning algorithms, and additional functions make it a powerful and flexible tool for advancing the use of machine learning in agriculture and related sectors.

Chapter 9

Future works

In addition to the findings of this study, several new lines of research can be pursued to further reinforce and improve these results. Firstly, the use of new clustering techniques can be explored to improve the efficiency of models during training. Future research could focus on temperature-based features beyond the DTR records used in this study, as well as the application of automated clustering algorithms to streamline the process of data analysis and model development.

Secondly, due to climatic aridity depends on precipitation, air temperature and reference evapotranspiration, the development of aridity index future maps could be a promising avenue for future research. One option could be using temperature and precipitation forecasts from the IPCC under different scenarios (such as AR5's RCP 4.5, RCP 6.0, and RCP 8.5). Thus, researchers could gain valuable insights into the future patterns of aridity, as well as the potential impact of climate change on these phenomena. Additionally, new developing scenarios like those presented in AR6 could be incorporated into this research to further enhance the accuracy and usefulness of future maps.

Lastly, the inclusion of new features that do not require in site instrumentation, such as using reanalysis grid datasets from Copernicus, could be explored to improve the accuracy and effectiveness of models. This approach would be particularly valuable for developing countries and areas where installing automatic weather stations (AWS) is not affordable, but accurate estimations of important parameters like ET_0, solar radiation, and aridity are crucial for agriculture and water resource management.