A bioinspired ensemble approach for multi-horizon reference evapotranspiration forecasting in Portugal

M. J. Jiménez-Navarro
Department of Computer Science, University of Seville
Seville, Spain
mjimenez3@us.es

M. Martínez-Ballesteros
Department of Computer Science, University of Seville
Seville, Spain
mariamartinez@us.es

I. Sofia Brito
Department of Engineering, Polytechnic Institute of Beja
Beja, Portugal
Instituto de Desenvolvimiento de Novas Tecnologias, Centre of Technology and Systems
Lisboa, Portugal
isabel.sofia@ipbeja.pt

F. Martínez-Álvarez
Data Science and Big Data Lab, Pablo de Olavide University
Seville, Spain
fmaralv@upo.es

G. Asencio-Cortés
Data Science and Big Data Lab, Pablo de Olavide University
Seville, Spain
guaasecor@upo.es

ABSTRACT

The year 2022 was the driest year in Portugal since 1931 with 97% of territory in severe drought. Water is especially important for the agricultural sector in Portugal, as it represents 78% total consumption according to the Water Footprint report published in 2010. Reference evapotranspiration is essential due to its importance in optimal irrigation planning that reduces water consumption. This study analyzes and proposes a framework to forecast daily reference evapotranspiration at eight stations in Portugal from 2012 to 2022 without relying on public meteorological forecasts. The data include meteorological data obtained from sensors included in the stations. The goal is to perform a multi-horizon forecasting of reference evapotranspiration using the multiple related covariates. The framework combines the data processing and the analysis of several state-of-the-art forecasting methods including classical, linear, tree-based, artificial neural network and ensembles. Then, an ensemble of all trained models is proposed using a recent bioinspired metaheuristic named Coronavirus Optimization Algorithm to weight the predictions. The results in terms of MAE and MSE are reported, indicating that our approach achieved a MAE of 0.658.

CCS CONCEPTS

• **Computing methodologies** → **Ensemble methods**; **Supervised learning by regression**; • **Applied computing** → **Environmental sciences**;

KEYWORDS

Time series, forecasting, bioinspired metaheuristic, evolutionary algorithm, agricultural, reference evapotranspiration, ensemble, deep learning

ACM Reference Format:
M. J. Jiménez-Navarro, M. Martínez-Ballesteros, I. Sofia Brito, F. Martínez-Álvarez, and G. Asencio-Cortés. 2023. A bioinspired ensemble approach for multi-horizon reference evapotranspiration forecasting in Portugal. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27-March 31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3555776.3578634

SAC '23, March 27-March 31, 2023, Tallinn, Estonia

https://doi.org/10.1145/3555776.3578634

1 INTRODUCTION

The *Drought in numbers* report [23] showed an increase of approximately 30% in the number of droughts worldwide since the year 2000. This has a social impact as people cannot access to potable water or food, as the agricultural and livestock sector cannot access to water to irrigate/feed their crops/cattle. An example is Portugal, where the 97% of the territory in severe drought was reported by the Instituto Português do Mar e da Atmosfera (IPMA) in 2022, which is the driest year since 1931. Drought is especially important in Portugal, as the agricultural sector represents 78% of total consumption according to the published 2010 report on the water footprint.

An essential parameter for the estimation of water resources is the reference evapotranspiration (ET_0), which is defined as the evapotranspiration of a hypothetical grass surface that is well watered. This parameter is used to estimate crop evapotranspiration (ET_c), which is mainly responsible for the irrigation infrastructure, scheduling, and management of a specific crop. An accurate estimate of ET_0 is essential, as an overestimation considerably increases the water footprint.

Machine learning and metaheuristics have been combined in multiple studies, obtaining astonishing results in fields like artificial vision, natural language processing, time series forecasting, etc.

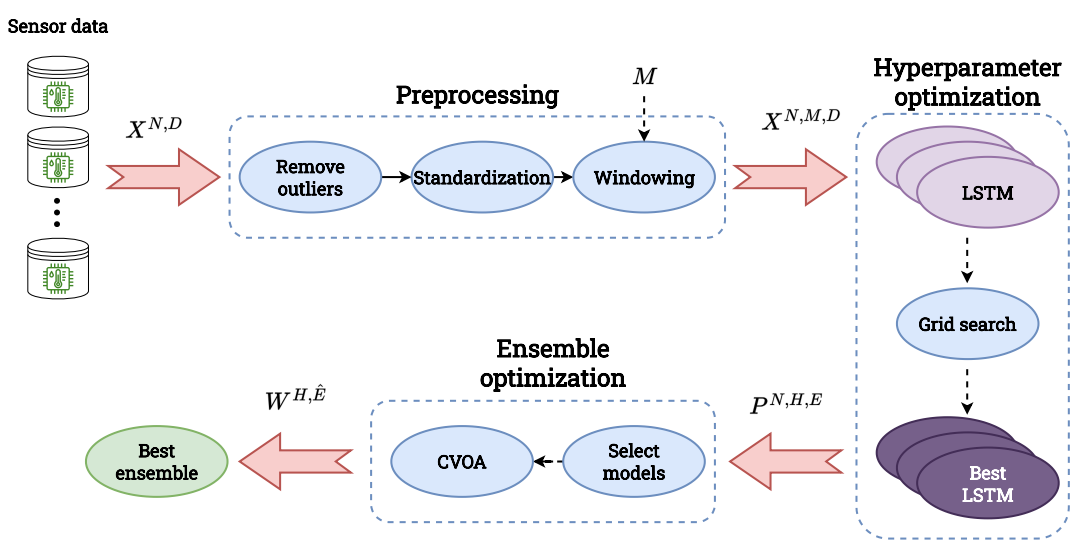

Figure 1: Complete methodology. The process starts collecting the data from the different stations into the matrix $X^{N,D}$ with N instances and D features. The first step preprocess the data, this step receives the hyperparameter M which influence in the windowing transformation and outputs the matrix $X^{N,M,D}$ with M past events. The second step finds the best hyperparameters for each selected model using a grid search. The third step selects the best combination of the models E using the predictions $P^{N,H,E}$ with H future instances and the CVOA algorithm. The CVOA algorithm obtains the optimal weights $W^{H,\hat{E}}$ for the best combination of models $\hat{E}$.

For this reason, this work proposes a framework for multi-horizon time series forecasting [1] that combines several state-of-the-art models using a bioinspired metaheuristic. Since the ensemble of different models generally obtains better results [14], we propose a set method based on the recently developed bioinspired metaheuristic named Coronavirus Optimization Algorithm (CVOA) [22], which has been shown to efficiently obtain good results in several scenarios [5, 28]. The role of the metaheuristic is to weigh the predictions produced by the different models. This framework is applied to a set of eight meteorological stations located near Beja (Portugal). Data contain multivariate records from 2012 to 2022 obtained from the Sistema Agrometeorológico para a Gestão da Rega no Alentejo (SAGRA) [10]. The goal is to forecast reference evapotranspiration without relying on public meteorological forecasts for three days in the future, as experts considered it enough to plan and manage water resources.

The main contributions of this paper are as follows:

(1) Development of an ensemble methodology with a bioinspired metaheuristic technique.
(2) Comparison between several state-of-the-art forecasting methods.
(3) Analysis of meteo-agricultural data and the impact of drought on the efficacy of the model.

The paper is structured as follows. Section 2 shows recent techniques used for evapotranspiration forecasting. Section 3 describes the methodology followed in the proposed framework step by step. Section 3.1 analyzes the dataset, its properties and the preprocessing applied. Section 4 details the experimental setting with all the information on the different forecasting methods applied, the evaluation metrics used, and the result comparison between the forecasting methods and the ensemble using our methodology. Finally, Section 5 shows the conclusions drawn and future work.

2 RELATED WORKS

There exist multiple studies that attempt to forecast reference evapotranspiration due to its multiple advantages.

R. Ballesteros et al. [4] use daily meteorological data to forecast six days ahead of the ET_0 calculated using the Penman–Monteith method in Spain. The meteorological data used as input are predictions produced by the Spanish Meteorological Service (AEMET). Hargreaves–Samani method is compared to an artificial neural network (ANN). The ANN generally achieves better results with a mean squared error of 0.98 mm/day. In our work, we present a comparison of several forecasting methods, including neural networks [19], demonstrating that an ensemble improves the results.

Y. Yang et al. [35] makes a seven days ahead forecast of ET_0 calculated with the Penman–Monteith method. The authors used daily meteorological data obtained from six stations and seven days ahead forecast in China. The method obtained a root mean square error equals to 0.98 mm/day, obtaining different behaviors at different stations. In our work, we do not depend on public meteorological forecasting, as there are contexts in which they are not available or adequately curated. We use different forecasting methods that use past data to make the calculation directly from the past meteorological data to the future ET_0.

L. B. Ferreira et al. [13] apply several models and forecasting strategies for a seven-day ET_0 forecast. The authors used data measured daily from 53 weather stations in Brazil, where 4 stations were

used for testing purposes. The forecasting methods include random forest, multilayer perceptron (MLP), long short-term memory (LSTM), convolutional neural network (CNN), and a CNN-LSTM combination. The forecasting strategies [21] were related to the output of the model: iterated forecasting, direct forecasting, and multiple input multiple output (MIMO). CNN-LSTM combination obtained the best results using the MIMO forecasting strategy with a root mean squared error of 0.87 mm/day. In our proposal, an ensemble approach is used to improve the results of the different models tested. Additionally, in addition to using the MIMO forecasting strategy, validation is done for a dataset subset of all stations, instead of all data in several stations as this scenario is more realistic.

P. de Oliveire e Lucas et al. [8] uses several CNN architectures and an ensemble of them to forecast daily ET_0 using the Penman–Monteith method ten days ahead. The authors used just the past information from ET_0 as input for the different models. Seasonal <utoregressive integrated moving average (SARIMA) and Seasonal Naive were used as a baseline, which is improved by the ensemble with a root mean square error of 0.94 mm/day. In our work, an ensemble is proposed without limiting them to a single type of model. Furthermore, the inclusion of the metaheuristic which improves the ensemble leveraging the relevance of the different models, is an important additional step included.

M. Alizamir et al. [3] use an adaptive neurofuzzy inference system (ANFIS) in combination with the genetic algorithm (GA) and particle swarm optimization (PSO) to make a one-day ahead forecast of ET_0, calculated using the Penman–Monteith method. The data was collected every month using two stations in Turkey, which contained several meteorological features. The proposed method was compared with a MLP and a decision tree. The proposed methods obtained an improvement of 27% compared to the best baseline model in terms of the root mean square error. In our study, we propose a multi-step forecasting model instead of using just one step ahead to analyze future behaviour in a short period.

A. Elbeltagi et al. [11] recollect monthly meteorolocal data from two stations in Pakistan to predict the ET_0 150 months ahead. The authors employ several tree-based models using bagging, voting, and random subspace ensemble methods to forecast ET_0. The different models in the ensemble are weighted using the Bayesian additive regression tree method [30]. The results conclude that the combination of additive regression with the M5' regression tree model [34] was the best combination with a root mean square error of 0.570. In our work, we use a more variate set of forecasting methods instead of using only tree-based methods, which increases the possibility of learning a more diverse representation of future behavior.

3 METHODOLOGY

The goal of the proposed methodology is to define a framework for time series forecasting using an ensemble of multiple heterogeneous models through the use of the CVOA bioinspired metaheuristic. This methodology is intended to be as efficient and scalable as possible, obtaining the best results using state-of-the-art models.

This section is structured as follows. Section 3.1 describes the dataset collected from different stations used to apply our methodology. Section 3.2 defines the preprocessing applied to the input data. Section 3.3 shows the different models used in the methodology. Section 3.4 describes the hyperparameter optimization process applied to the different models to find the best hyperparameter configuration. Section 3.5 details the bioinspired ensemble process used to combine and improve the different models. Figure 1 shows the complete methodology proposed in this work, which is described in the following sections.

3.1 Dataset

The dataset studied contains data collected from eight automatic meteorological stations on a daily basis in Beja (Portugal). These data are part of SAGRA of the Centro Operativo e de Tecnologia de Regadio [10], which has recollected data from 2012 to 2022. In total, the dataset contains nearly 3525 instances for each station (eight in total).

Each station has been named according to its respective zone; the names are as follows: Castro Verde (CV), Estremoz (E), Herdade Lameiroes (HL), Herdade Outeiro (HO), Quinta Saude (QS), Serpa (S), Viana Alentejo (VA) and Vidigueira (V). The sensors at the stations collect 14 meteorological data and the reference evapotranspiration using the Penman-Monteith method.

Feature	Full name	Mean	STD	Min	Max
Tmed (°C)	Mean temperature	16.69	6.22	1	35.16
Tmax (°C)	Maximum temperature	24.45	7.89	2.38	46.52
Tmin (°C)	Minimum temperature	9.87	5.07	-5.37	24.98
HRmed (%)	Mean relative humidity	70.85	16.10	22.48	102.52
HRmax (%)	Maximum relative humidity	93.82	7.67	65.19	100
HRmin (%)	Minimum relative humidity	42.62	18.82	-0.4	100
RSG (kj/m2)	Global Solar Radiation	17083.96	8087.03	641.84	35993
DV (graus)	Wind direction	228.75	100.29	0	360
VVmed (m/s)	Mean wind speed	1.67	0.85	0	4.53
VVmax (m/s)	Maximum wind speed	6.28	2.36	0	116.05
P (mm)	Precipitation	1.31	4.60	0	88.4
Tmed Relva (°C)	Mean soil temperature	18.55	7.72	1.4	40.29
Tmax Relva (°C)	Maximum soil temperature	26.29	12.69	5.83	71.43
Tmin Relva (°C)	Minimum soil temperature	13.51	5.86	-4.8	30.7
ET_0 (mm)	Reference evapotranspiration	3.43	2.01	0	9.47

Table 1: Statistics of features collected at each station.

Table 1 shows the different features obtained from each station, their abbreviations, and some basic statistics. In general terms, the ambient temperature and the soil temperature seem to remain in optimal ranges ([15 °C, 30 °C] depending of the crop) to ensure optimal growth [16, 26]. Humidity is generally high between a range of [42 %, 94%]. The wind seems to be moderate in general, with no dangerous values. Precipitations are mainly scarce, which implies the use of external water sources.

Figure 2 shows the evolution of each feature over time. We can observe two types of feature: seasonal and mostly chaotic. Chaotic variables such as maximum relative humidity, wind measurements, and precipitations do not present a clear temporal pattern. Seasonal features show a temporal pattern in which the warm seasons present the peaks in temperature features and valleys in humidity features. ET_0 seems to present a pattern similar to temperature features.

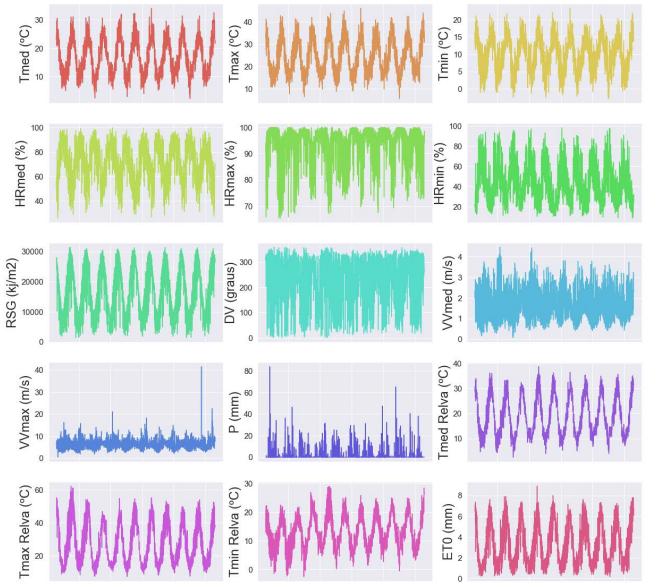

Figure 2: Evolution of the features over time.

3.2 Data preprocessing

The section describes the processing applied to the dataset to prepare the data for the different architectures [2, 27]. Processing is divided into three steps: outlier processing, standardization, division, and windowing. The process is executed sequentially and receives the input $X^{N,D}$ with N instances ordered by station and date, corresponding to the union of all station instances, and D features, corresponding to the 15 features measured at each station.

The dataset presents some extreme outliers in some features, mostly in the mean, maximum, and minimum grass temperature. The absolute value of the z-score has been calculated for each instance in every feature to identify these outliers. Once the z-score is calculated, instances with a z-score greater than three are treated as outliers and removed. Then, the removed instances are imputed using the K nearest neighbors (KNN) algorithm with $K = 3$.

Once the outliers have been removed and imputed for each feature, a standardization process has been applied to set the mean to zero and the standard deviation to 1. For the same reason as for the imputation, just one mean and one standard deviation were calculated using all the station data. Therefore, the same mean and standard deviation were used to characterize all stations.

The data contain instances that have been measured for almost 10 years in total. The division has been applied using instances over a full year. Thus, the train data contain data from 2012 to 2019 included, the validation data use the year 2020 and the test data contain the year 2021 and 2022 included. Usually, only the last year is included in the test. However, since 2022 has only 8 months and does not complete a period, 2021 was also included. Furthermore, an analysis to see the effect of drought and compare it with the performance obtained in 2021 will be presented in Section 4.

The last step is the windowing process applied to every division, which transforms the two-dimensional matrix with every instance as a row and every feature as a column in a three-dimensional tensor. The new dimension includes the past information of an instance, feeding the architectures with temporal information. This means that an instance X_t^D with D features at a moment t obtains new information from M past instances with respect to the moment t. This means that each instance now contains the features X_t^D and the features $X_{t-M-1}^D \cdots X_{t-1}^D$. Thus, the entire dataset and the output of the preprocessing step compose a three-dimensional matrix X^{NxMxD} with N instances ordered by station and date, M past information, and D features. Note that, since the matrix $X^{N,D}$ contains data from different stations, it is necessary to remove each window that mixes information from different stations.

3.3 Forecast strategy

The input of the architectures is a set of meteorological data obtained from multiple stations. In this work, we propose a one-to-many approach (one model for many stations) to produce the forecasts for the different stations. This approach is commonly used in grouped datasets using the bottom-up strategy [24, 29]. This approach uses one model to produce the forecast for all stations rather than using a model for each station. This produces a more scalable methodology which requires less resources as less information needs to be stored and use less computation for retraining the model. Additionally, the input of each model must include information from one station in order to achieve better scalability. Using information from multiple stations can produce good results, but this approach limits stations to centralized information sharing. Another problem can arise if a new station needs to be included in the model; this may change the entire architecture and would require one to retrain from the beginning. Using a single model with the input of a single station each time opens up the possibility of using transfer learning over another new station or a problem of similar nature.

For the output of the models, a direct approach [17] has been applied to obtain the three steps ahead forecasting as output. For example, a multiple input corresponding to the windowed meteorological features is fed into the models, and three outputs are obtained.

3.4 Hyperparameter optimization

In this step, the comparison of 14 different commonly used models was used using the same window sizes and forecasting horizons using the same random seed to ensure that randomness has the smallest possible influence. In this step, a grid search [33] has been applied to perform a homogeneous search throughout the hyperparameter space, reducing the effect of randomness in comparison.

Table 2 shows the different models used to evaluate the dataset and the hyperparameter space for each one. The hyperparameters have been selected to cover the most common parametrization trying to give equal opportunities to each model.

In the case of neural networks (MLP, CNN, LSTM), the selected optimizer was Adam, and the activation function was Relu except in LSTM as this configuration has proved to provide competitive results in several applications. The hyperparameters of the windowing process are optimized jointly with the hyperparameters of each model using a range from 3 to 7 past instances. A linear

Model	Hyperparameter space
Lasso [31]	{'Alpha': [0.0001, 0.001, 0.01, 0.1, 1, 2, 3]}
Ridge [32]	{'Alpha': [0.0001, 0.001, 0.01, 0.1, 1, 2, 3]}
Linear	-
ElasticNet [36]	{'Alpha': [0.0001, 0.001, 0.01, 0.1, 1, 2, 3], 'Ratio': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]}
Decision tree [25]	{'Splitter': ['best', 'random'], 'Depth': [1, 3, 5, 7, 9], 'Criterion': ['MSE', 'Friedman MSE', 'MAE']}
SVM [7]	{'C': [0, 1, 2, 3, 4], 'Kernel': ['Linear', 'Polynomial', 'RBF', 'Sigmoid']}
MLP [12]	{'Layers': [1,2,3], 'Units': [4, 8, 16]}
CNN [12]	{'Layers': [1,2,3], 'Units': [4, 8, 16], 'Kernel size': [3]}
LSTM [12]	{'Layers': [1,2,3], 'Units': [4, 8, 16]}
Random Forest [15]	{'Estimators': [50, 100, 200, 300, 400, 500], 'Depth': [1, 3, 5, 7, 9], 'Criterion': ['MSE', 'Friedman MSE', 'MAE']}
XGB [6]	{'Estimators': [50, 100, 200, 300, 400, 500], 'Depth': [2, 3, 4, 6, 7], 'Criterion': ['MSE'], 'ETA': [0.3, 0.4, 0.5]}
CatBoost [9]	{'Depth': [1, 3, 5, 7, 9], 'Learning rate': [0.003, 0.001, 0.0001], 'Iterations': [50, 100, 200, 300, 400, 500], 'L2': [0, 1, 2, 3, 4]}
LGBM [20]	{'Estimators': [50, 100, 200, 300, 400, 500], 'Depth': [1, 3, 5, 7, 9], 'Learning rate': [0.003, 0.001, 0.0001], 'Regularization': [0, 0.01, 0.03]}

Table 2: Model and hyperparameters used in the grid search.

model is also used in the grid search, which is not included because it does not have hyperparameters.

3.5 Ensemble

In this step, the predictions are improved by combining several models to reduce bias and improve efficacy. The ensemble process is divided into three steps. The first finds the selection of a combination of the best E models obtained from Section 3.4. The second step obtains the predictions $P^{V,H,E}$ of the selected models E and the H future (ET_0) prediction in all instances of the train and validation sets V. Finally, the CVOA algorithm is used to determine the weights $W^{H,E}$ for the predictions of the different models E and their horizons H. Initially, the weights for all models and horizons are the same using a base voting approach. Individuals in the CVOA algorithm consist of a matrix of size HxE which, initially, is unrestricted. This means that the weights for all models E for a specific horizon may not sum up 1. For that reason, a normalization for each row in H is applied to make the weights in the combination of E models sum up 1.

Therefore, the ensemble consists of two search algorithms. The first search algorithm must select a combination of models from which the second search obtains the weights of the predictions for these models. CVOA algorithm has a set of individuals that represent the search space and, potentially, can be infected by evaluating its performance. As the second search algorithm, CVOA has proven to be a very efficient algorithm that can converge in a few iterations, we propose a grid search over every pair, trio, and quartet combination of models. This will allow us to find the best combination of models based on CVOA weights.

The ensemble allows us to use heterogeneous models that require different window sizes and hyperparameters. Furthermore, this ensemble method is efficient, as it does not introduce additional computation at inference time.

4 RESULTS

4.1 Evaluation metrics

To evaluate the efficacy and efficiency of the different models, three common metrics have been selected. For efficacy, mean absolute error (MAE) and mean squared error (MSE) were selected because they are well-known metrics that allow us to obtain a good interpretation of the results. Additionally, the weighted average percentage error (WAPE) metric is included to avoid unrepresentative values shown when the values are below zero, as in our study. As we are in a multi-step forecasting, the metrics apply the calculations averaging the results obtained by each horizon. Taking into account N the number of instances to be evaluated, H the forecast horizon, $y(t)_{t+h}$ and $\hat{y}(t)_{t+h}$ the true and predicted value correspondingly at moment t and future prediction at time $t+h$, the formulas applied are the following:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{H}\sum_{h=1}^{H}|y(t)_{t+h} - \hat{y}(t)_{t+h}| \tag{1}$$

$$MSE = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{H}\sum_{h=1}^{H}(y(t)_{t+h} - \hat{y}(t)_{t+h})^2 \tag{2}$$

$$WAPE = \frac{1}{H}\sum_{h=1}^{H}\frac{\sum_{t=1}^{N}|y(t)_{t+h} - \hat{y}(t)_{t+h}|}{\sum_{t=1}^{N}|y(t)_{t+h}|} \tag{3}$$

For efficiency, the training time measured in minutes is selected as represents the metric that takes more resources and is one of the most important metrics in order to allow the model to be re-train. The hardware configuration consists of an NVIDIA Geforce RTX 3070 GPU, an AMD Ryzen 7 5800X 3.8 GHz, and 32 GB of RAM. Note that the GPU is used in CNN, LSTM, Extreme Gradient Boosting (XGB) and Light Gradient Boosted Machine (LGBM) models.

4.2 Hyperparameter optimization results

This section represents the best results obtained for each model after applying the grid search. The MSE was used as a criterion to decide which hyperparameters are the best. The data was sorted by MSE reporting efficacy metrics and efficiency. Additionally, a baseline model (Base) was included consisting of repeating the last known value for the three-days forecasting, this approach has a training time of zero as there is no model implied.

Table 3 shows the efficacy and efficiency metrics for the best models found in the grid search. The results have been ordered by MSE, showing that the best model is the LSTM with a training time fewer than that of almost all models except the linear model. XGB and Random forest show efficacy similar to the best model, but considering efficiency, LSTM shows a 3x and 8x reduction in training time. However, it must be considered that LSTM and XGB used a GPU for the training process, while random forest did not. The next eight models show similar efficacy in general with heterogeneous efficiency. SVM shows the worst efficiency in terms

Model	MAE	MSE	WAPE	Time (m)
LSTM	0.669	0.774	0.183	0.272
XGB	0.678	0.784	0.186	1.014
Random Forest	0.673	0.788	0.184	7.032
MLP	0.692	0.811	0.190	18.860
CNN	0.691	0.812	0.189	0.514
Lasso	0.700	0.827	0.191	0.562
ElasticNet	0.701	0.828	0.191	0.783
SVM	0.688	0.832	0.187	762.736
Ridge	0.704	0.833	0.192	0.062
Linear	0.704	0.833	0.192	0.100
Decision Tree	0.706	0.889	0.193	0.498
LGBM	0.762	0.927	0.208	4.691
CatBoost	0.790	0.987	0.215	168.259
Base	0.855	1.350	0.234	-

Table 3: Top models obtained after the grid search sorted by MSE.

of training time in all the datasets with a great difference followed by CatBoost and MLP. Interestingly, neural network approaches are in the top five best models, while only two of four ensemble methods obtained good results. The other two ensemble models obtained the worst efficacy, being just better than the baseline model.

Model	Window size	Optimal parameters
LSTM	3	{'Units': 16, 'Layers': 3}
XGB	4	{'ETA': 0.3, 'Criterion': 'MSE', 'Depth': 3, 'Estimators': 50}
Random forest	4	{'Criterion': 'Friedman MSE', 'Depth': 9, 'Estimators': 200}
MLP	3	{'Layers': 3, 'Units': 4}
CNN	4	{'Units': 4, 'Layers': 2, 'Kernel size': 3}
Lasso	7	{'Alpha': 0.001}
ElasticNet	7	{'Alpha': 0.001, 'Ratio': 0.7}
SVM	7	{'C': 1, 'Kernel': 'Linear'}
Ridge	7	{'Alpha': 3}
Linear	7	-
Decision tree	4	{'Criterion': 'Friedman MSE', 'Depth': 7, 'Splitter': 'best'}
LGBM	5	{'Learning rate': 0.003, 'Depth': 7, 'Estimators': 500, 'Regularization': 0}
CatBoost	7	{'Depth': 9, 'Iterations': 500, 'Regularization': 0, 'Learning rate': 0.003}

Table 4: Best parameters obtained by the grid search.

Table 4 shows the best parameters found by grid search using the same order as Table 3. Focusing on the optimal hyperparameters, there seem to be some extreme scenarios in some cases. For example, the best parameters of the LSTM were the highest number of units and layers, and, for XGB, the smallest depth and the lowest number of estimators. LGBM and CatBoost also obtained a greater number of depths and estimators/iterations, which could be related to their poor efficacy in producing overfitting.

Considering the window size, a value of 3 or 4 seems to be the most common value in the top 5 models. Interestingly, all linear models required the maximum window size. These facts lead us to believe that using nonlinear models like neural networks [18] or tree-based models requires fewer past events than linear models, except LGBM and CatBoost.

4.3 Ensemble results

This section describes the results obtained by the ensemble process detailed in the methodology. This section is structured as follows. Section 4.3.1 shows the top ten combinations that improve the best model obtained during the grid search. Section 4.3.2 represents the error obtained in the best ensemble for each station and horizon. Section 4.3.3 compares the results obtained by the best ensemble for the same periods in 2021 and 2022 for each station. Section 4.3.4 analyzes the results obtained by the best model for each station and season. Finally, Section 4.3.5 represents the weights learned by CVOA for the best ensemble.

Models	MAE	MSE	WAPE	Time (m)
LSTM, XGB, CNN	**0.658**	**0.747**	**0.180**	**1.80**
LSTM, XGB, MLP, CNN	0.658	0.747	0.180	20.66
LSTM, XGB, MLP, LGBM	0.661	0.748	0.180	24.84
LSTM, XGB, MLP, ElasticNet, LGBM	0.662	0.748	0.180	25.62
LSTM, XGB, MLP	0.659	0.748	0.180	20.15
LSTM, XGB, LGBM	0.660	0.748	0.180	5.98
LSTM, XGB, MLP, CNN, Random Forest	0.659	0.749	0.180	27.69
LSTM, XGB	0.659	0.749	0.180	1.29
LSTM, XGB, MLP, CNN, LGBM	0.661	0.749	0.181	25.35
LSTM, XGB, MLP, Lasso, ElasticNet, LGBM	0.663	0.749	0.181	26.19

Table 5: Top ten combinations obtained from the ensemble stage sorted by MSE.

4.3.1 Combination analysis. Table 5 shows the top ten combinations ordered by MSE. The results show that the best ensemble consists of a combination of LSTM, XGB, MLP, and CNN, with an improvement over LSTM of 3.50% in MSE with an increase of 75 times the training time. Improvement in efficacy is acceptable, however, the MLP introduced too much increment of training time. For this reason, we consider the best ensemble to be the combination of LSTM, XGB, and CNN (marked in bold) that has the same improvement in efficacy (3.5%) over LSTM and the efficiency increases almost 6 times.

From Table 5 we can analyze the most effective models. We observed that the combination of LSTM and XGB seems to be mandatory to obtain good results, since both models appear in all ensembles. In fact, using only both models, the results show that there is an improvement of 3.2% over LSTM in terms of MSE and the training time increases just almost 4 times.

Models like LGBM, MLP, Random Forest, Lasso, and ElasticNet seem to be useful too, but introduce great training time increments. Despite the fact that the LGBM model obtains poor results during Section 3.4, it appears to slightly improve the efficacy. Furthermore, models like Decision Tree, Ridge, SVM, and CatBoost seem to be useless in combination with other models.

	MAE			MSE			WAPE		
	1	2	3	1	2	3	1	2	3
CV	0.626	0.738	0.796	0.676	0.915	1.062	0.150	0.177	0.190
E	0.566	0.652	0.694	0.566	0.710	0.793	0.176	0.205	0.218
HL	0.494	0.596	0.619	0.444	0.593	0.638	0.145	0.177	0.186
HO	0.551	0.646	0.677	0.550	0.712	0.787	0.153	0.179	0.186
QS	0.607	0.716	0.754	0.632	0.836	0.917	0.159	0.188	0.198
S	0.590	0.690	0.711	0.648	0.811	0.859	0.156	0.184	0.192
V	0.567	0.689	0.719	0.584	0.822	0.902	0.156	0.191	0.200
VA	0.592	0.723	0.774	0.622	0.866	0.978	0.160	0.196	0.210

Table 6: MAE, MSE, and WAPE error analysis by station and horizon.

4.3.2 Horizon analysis. Table 6 shows the error obtained for the first, second, and third horizons independently for each station.

In general, the error increases as the model forecast further ahead in time, obtaining an increase of almost 33% between the first and second horizons and 10% between the second and third horizons on average.

There exist significant differences between the stations. The station with a lower error corresponds to Herdade Lameiroes, while the station with a higher error corresponds to Castro Verde.

4.3.3 Year analysis. To analyze the effect of drought and the possible drift of the distribution, a comparison between 2021 and 2022 is presented. However, since the year 2022 only includes eight months, only the first eight months of 2021 have been considered in this comparison.

	MAE		MSE		WAPE	
	2021	2022	2021	2022	2021	2022
CV	0.733	0.798	0.878	1.041	0.169	0.180
E	0.708	0.681	0.824	0.756	0.204	0.208
HL	0.599	0.648	0.611	0.658	0.163	0.184
HO	0.638	0.702	0.712	0.838	0.166	0.182
QS	0.709	0.785	0.814	0.994	0.173	0.198
S	0.691	0.763	0.798	0.983	0.172	0.190
V	0.702	0.746	0.847	0.939	0.180	0.195
VA	0.721	0.753	0.889	0.909	0.184	0.195

Table 7: MAE, MSE and WAPE comparison between the first 8 months of 2021 and 2022 for each state.

Table 7 shows the comparison by year for each station. The stations perform similarly as reported in the previous section, Herdade Lameiroes being the best station and Castro Verde the worst station. All stations considerably increased the error during 2022, which may be caused by the effect of the distribution drift caused by the drought. Estremoz is the only station that improves the results during 2022, this may be explained because the drought does not have a high impact on this station.

4.3.4 Season analysis. As shown in Section 3.1, the warm seasons have different behaviours compared to the cold seasons. For that reason, it is interesting to analyze the effect of seasons on the error produced by the selected ensemble.

	MSE				WAPE			
	Fall	Spring	Summer	Winter	Fall	Spring	Summer	Winter
CV	0.401	1.260	1.084	0.575	0.185	0.181	0.118	0.281
E	0.325	1.121	0.504	0.559	0.220	0.227	0.104	0.319
HL	0.206	0.979	0.436	0.383	0.172	0.194	0.090	0.276
HO	0.298	1.014	0.663	0.547	0.187	0.177	0.107	0.286
QS	0.418	1.176	0.654	0.686	0.206	0.188	0.105	0.303
S	0.268	1.284	0.605	0.612	0.175	0.188	0.106	0.292
V	0.276	1.305	0.646	0.547	0.185	0.198	0.108	0.296
VA	0.438	1.272	0.855	0.515	0.211	0.203	0.120	0.292

Table 8: MSE and WAPE error analysis by season and station.

Table 8 shows the error obtained for each season and station. As observed, the major error occurs during the spring season, which is the season in which the flowering process begins and the temperature begins to increase. Interestingly, Castro Verde is not the worst station during Spring but Vidigueira which increases considerably the error compared with other seasons. The greater error usually found in Castro Verde is explained because, in contrast to other stations, the error remains considerably high during the summer, while the error in other stations decreases. Fall is the season with the lowest error, in general, where Herdade Lameiroes has the lowest error and Viana Alentejo has the greatest one. Winter and summer are generally similar, and summer has more errors, in general.

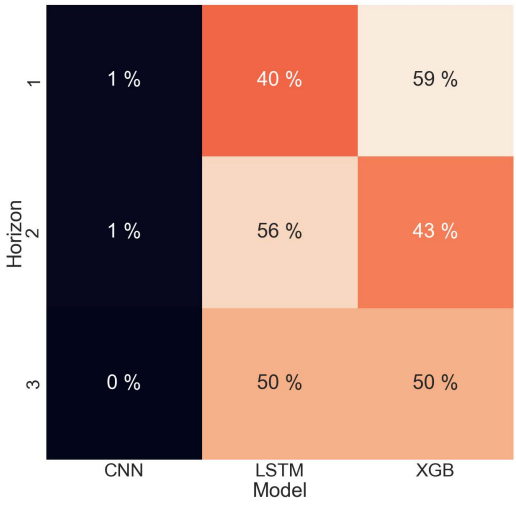

Figure 3: Weight matrix learned by the CVOA algorithm.

4.3.5 Weight analysis. Figure 3 represents the weights learned by CVOA that assign the relevance of each model and horizon. The weights are represented as a matrix in which the rows represent the horizons and the columns represent the combined models. Note that each row must sum 1 to produce the final prediction.

The matrix shows that LSTM and XGB are the most relevant models in all horizons. In the first horizon, XGB seems to be the

most relevant, while in the second horizon, LSTM seems to be more relevant than XGB. In the third horizon, LSTM and XGB are equally relevant. As observed, CNN seems to have little or no relevance, slightly reducing the error with litter contribution.

5 CONCLUSIONS AND FUTURE WORKS

In this work, we propose a general comparison between different models and the bioinspired ensemble of them applied to the prediction of reference ET_0 forecasting using the CVOA algorithm. The results showed that the ensemble of boosting (XGB) and deep learning models (LSTM and CNN) are the best individually in terms of efficacy with a remarkable improvement. Specifically, LSTM and XGB were the most important models that reduced the impact of bias for the ET_0 forecasting.

In future work, other forecasting strategies, such as a model for each station, may be tested and compared between them, selecting the best strategy which adapts to the business requirements. Feature engineering may be used, as the effectiveness of deep learning could be related to its ability to extract features automatically.

ACKNOWLEDGMENTS

The authors would like to thank the Centro Operativo e de Tecnologia de Regadio (COTR), the Portuguese Agency *Fundação para a Ciência e a Tecnologia* (FCT) in the framework of the project UIDB/00066/2020, the Spanish Ministry of Science and Innovation for the support under the project PID2020-117954RB, the European Regional Development Fund and *Junta de Andalucía* for projects PY20-00870 and UPO-138516.

REFERENCES

[1] A. Galicia and R. Talavera-Llames and A. Troncoso and I. Koprinska and F. Martínez-Álvarez. 2019. Multi-step forecasting for big data time series based on ensemble learning. *Knowledge-Based Systems* 163 (2019), 830–841.

[2] A. Agathiyan, M. Naresh, and M. Immanuel. 2021. Personalized Weather Station With Machine Learning Based Rainfall Predictor. *International Journal of Advanced Research in Science, Communication and Technology* 4 (2021), 710–715.

[3] M. Alizamir, O. Kisi, R. M. Adnan, and A. Kuriqi. 2020. Modelling reference evapotranspiration by combining neuro-fuzzy and evolutionary strategies. *Acta Geophysica* 68, 4 (2020), 1113–1126.

[4] R. Ballesteros, J. F. Ortega, and M. A. Moreno. 2016. FORETo: New software for reference evapotranspiration forecasting. *Journal of Arid Environments* 124 (2016), 128–141.

[5] K. T. Bui, J. F. Torres, D. Gutiérrez-Avilés, V. Nhu, D. T. Bui, and F. Martínez-Álvarez. 2022. Deformation forecasting of a hydropower dam by hybridizing a long short-term memory deep learning network with the coronavirus optimization algorithm. *Computer-Aided Civil and Infrastructure Engineering* 37, 11 (2022), 1368–1386.

[6] T. Chen and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (ACM/SIGKDD)*. New York, NY, USA, 785–794.

[7] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[8] P. de Oliveira e Lucas, M. A. Alves, P. C. de Lima e Silva, and F. G. Guimarães. 2020. Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks. *Computers and Electronics in Agriculture* 177 (2020), 105700.

[9] A. V. Dorogush, V. Ershov, and A. Gulin. 2018. CatBoost: gradient boosting with categorical features support.

[10] Centro Operativo e de Tecnologia de Regadio. 2022. Sistema Agrometeorológico para a Gestão da Rega no Alentejo. http://www.cotr.pt/servicos/sagra.php.

[11] A. Elbeltagi, A. Raza, Y. Hu, N. Al-Ansari, N. Kushwaha, A. Srivastava, D. Kumar Vishwakarma, and M. Zubair. 2022. Data intelligence and hybrid metaheuristic algorithms-based estimation of reference evapotranspiration. *Applied Water Science* 12, 7 (2022), 1–18.

[12] J. Torres F, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso. 2021. Deep learning for time series forecasting: a survey. *Big Data* 9, 1 (2021), 3–21.

[13] L. B. Ferreira and F. F. da Cunha. 2020. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. *Computers and Electronics in Agriculture* 178 (2020), 105728.

[14] D. Hadjout, J. F. Torres, A. Troncoso, A. Sebaa, and F. Martínez-Álvarez. 2022. Electricity consumption forecasting based on ensemble deep learning with application to the Algerian market. *Energy* 243 (2022), 123060.

[15] T. K. Ho. 1995. Random decision forests. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. Montreal, QC, Canada, 278–282.

[16] R. Huang, J. Huang, C. Zhang, H. Ma, W. Zhuo, Y. Chen, D. Zhu, Q. Wu, and L. R. Mansaray. 2020. Soil temperature estimation at different depths, using remotely-sensed data. *Journal of Integrative Agriculture* 19, 1 (2020), 277–290.

[17] Y. In and J. Jung. 2021. Simple averaging of direct and recursive forecasts via partial pooling using machine learning. *International Journal of Forecasting* 2021 (2021), 1–14.

[18] M. J. Jiménez-Navarro, M. Martínez-Ballesteros, I. S. Sousa Brito, F. Martínez-Álvarez, and G. Asencio-Cortés. 2023. Feature-Aware Drop Layer (FADL): A Nonparametric Neural Network Layer for Feature Selection. In *Proceedings of International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)*. Salamanca, Spain, 557–566.

[19] M. J. Jiménez-Navarro, G. Asencio-Cortés, A. Troncoso, and F. Martínez-Álvarez. 2021. HLNet: A Novel Hierarchical Deep Neural Network for Time Series Forecasting. In *Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)*. Bilbao, Spain, 721–727.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30 (2017), 3146–3154.

[21] B. Lim and S. Zohren. 2021. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A* 379, 2194 (2021), 20200209.

[22] F. Martínez-Álvarez, G. Asencio-Cortés, J. F. Torres, D. Gutiérrez-Avilés, L. Melgar-García, R. Pérez-Chacón, C. Rubio-Escudero, J. C. Riquelme, and A. Troncoso. 2020. Coronavirus Optimization Algorithm: a bioinspired metaheuristic based on the COVID-19 propagation model. *Big Data* 8, 4 (2020), 308–322.

[23] United Nations. 2022. Drought in numbers. https://www.unccd.int/resources/publications/drought-numbers.

[24] O. O. Owolabi and D. A. Sunter. 2022. Bayesian Optimization and Hierarchical Forecasting of Non-Weather-Related Electric Power Outages. *Energies* 15, 6 (2022), 1–22.

[25] R. J. Quinlan. 1992. Learning with Continuous Classes. In *Proceedings of Australian Joint Conference on Artificial Intelligence (AJCAI)*. Singapore, 343–348.

[26] H. Sanikhani, R. C. Deo, Z. M. Yaseen, O. Eray, and O. Kisi. 2018. Non-tuned data intelligent model for soil temperature estimation: A new approach. *Geoderma* 330 (2018), 52–64.

[27] M. Sattar, A. Najah, N. Zaini, A. Razzaq, P. Kumar, M. Sherif, A. Sefelnasr, and A. El-Shafie. 2021. Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia. *Scientific Reports* 11 (2021), 18935.

[28] C. Segarra-Martín, M. Martínez-Ballesteros, A. Troncoso, and F. Martínez-Álvarez. 2022. A Novel Approach to Discover Numerical Association Based on the Coronavirus Optimization Algorithm. In *Proceedings of Symposium on Applied Computing (ACM/SIGAPP)*. New York, NY, USA, 1148–1151.

[29] M. Soto-Ferrari, O. Chams-Anturi, J. P. Escorcia Caballero, N. Hussain, and M. Khan. 2019. *Evaluation of Bottom-Up and Top-Down Strategies for Aggregated Forecasts: State Space Models and ARIMA Applications*. Enschede, The Netherlands, 413–427.

[30] R. Sparapani, C. Spanbauer, and R. McCulloch. 2021. Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software* 97, 1 (2021), 1–66.

[31] R. Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)* 58 (1996), 267–288.

[32] A. N. Tikhonov. 1963. Solution of Incorrectly Formulated Problems and the Regularization Method. *Soviet Mathematics Doklady* 4 (1963), 1624–1627.

[33] J. Torres, A. Galicia de Castro, A. Troncoso, and F. Martínez-Álvarez. 2018. A scalable approach based on deep learning for big data time series forecasting. *Integrated Computer-Aided Engineering* 25 (2018), 1–14.

[34] Y. Wang and I. H. Witten. 1997. Inducing Model Trees for Continuous Classes. In *Proceedings of European Conference on Machine Learning Poster Papers (ECML)*. Hamilton, New Zealand, 128–137.

[35] Y. Yang, Y. Cui, Y. Luo, X. Lyu, S. Traore, S. Khan, and W. Wang. 2016. Short-term forecasting of daily reference evapotranspiration using the Penman-Monteith model and public weather forecasts. *Agricultural Water Management* 177 (2016), 329–339.

[36] H. Zou and T. Hastie. 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B* 67 (2005), 301–320.