Applied Intelligence (2023) 53:24765–24781
https://doi.org/10.1007/s10489-023-04783-2

Evaluation of synthetic data generation for intelligent climate control in greenhouses

Juan Morales-García[1] · Andrés Bueno-Crespo[1] · Fernando Terroso-Sáenz[1] · Francisco Arcas-Túnez[1] · Raquel Martínez-España[2] · José M. Cecilia[3]

Accepted: 11 June 2023 / Published online: 28 July 2023

Abstract

We are witnessing the digitalization era, where artificial intelligence (AI)/machine learning (ML) models are mandatory to transform this data deluge into actionable information. However, these models require large, high-quality datasets to predict high reliability/accuracy. Even with the maturity of Internet of Things (IoT) systems, there are still numerous scenarios where there is not enough quantity and quality of data to successfully develop AI/ML-based applications that can meet market expectations. One such scenario is precision agriculture, where operational data generation is costly and unreliable due to the extreme and remote conditions of numerous crops. In this paper, we investigated the generation of synthetic data as a method to improve predictions of AI/ML models in precision agriculture. We used generative adversarial networks (GANs) to generate synthetic temperature data for a greenhouse located in Murcia (Spain). The results reveal that the use of synthetic data significantly improves the accuracy of the AI/ML models targeted compared to using only ground truth data.

Keywords Deep learning · Synthetic time series data generation · Generative adversarial networks · Time series forecasting

1 Introduction

Modern technologies provide sustainable and feasible solutions to many real-world problems. One area where these technologies have provided solutions in recent years is agriculture. Precision agriculture applies innovative technologies to the agricultural world to reduce costs, increase profit and achieve sustainability [1]. A comprehensive review of the state of the art use of artificial intelligence (AI) in smart greenhouses is provided by [2]. This review focused on the optimization of crop yields, reduction of water consumption, fertilizers, diseases, pests, and the search for improved agricultural sustainability. Therefore, the status of various AI technologies in smart greenhouses is reviewed by discussing the extent to which technologies have been successfully applied in an agricultural context and the options for optimizing their usability.

Among the challenges facing precision agriculture is the adaptation of processes to climate change [3]. To monitor crop status to face sudden weather changes that occur mainly in semi-arid climates, farmers use technologies such as the Internet of Things (IoT) to monitor their plots and/or greenhouses [4, 5]. The data generated by these systems also feed into decision support systems to perform intelligent and automatic actions on the plots. Several leading examples for these include climate control in greenhouses [6] or frost prevention in a fruit orchard through smart irrigation [7].

Although decision support systems have numerous advantages and can make decisions in anticipation of future climatic conditions, they have the disadvantage of needing to create local models to achieve high accuracy in predicting climate variables [8, 9]. This disadvantage translates into the need to have historical data on the location of the plot to train and create an accurate model according to the farmer's needs. This would mean installing the IoT system to collect data but not accurately using the prediction system until there is sufficient historical data to create the prediction model. In [10], the authors review four bio-inspired intelligent algorithms used for agricultural applications, such as ecological, swarm intelligence-based, ecology-based, and multi-objective-based algorithms. Some observed that no universal algorithm could perform multiple functions on farms; therefore, different algorithms were designed according to the specific functions to be performed.

✉ Juan Morales-García
jmorales8@ucam.edu

Extended author information available on the last page of the article

Despite being in the era of Big Data, there is still a lack of quality data to address local problems such as the one mentioned above [11]. Recently, AI techniques have emerged that can generate artificial data of equal or higher quality than the original data, thus solving the problem of the amount of data needed to train local models [12]. Among these techniques, generative adversarial networks (GANs) – deep artificial neural networks capable of generating artificial data – [13] have obtained interesting results in different applications, including image processing [14], speech recognition [15] and other [16].

Within the field of precision agriculture, GANs have recently been applied to image processing tasks such as image augmentation [17, 18] and other tasks within computer vision [19]. However, to the best of our knowledge, synthetic data generation has not been applied to time series data generation in precision agriculture for climate control. In this study, we propose and evaluate synthetic data generation strategies to increase the accuracy of forecasting models for greenhouse climate control.

Greenhouses are agricultural structures that must be tightly controlled to avoid extreme weather conditions to achieve high crop yields [20]. Therefore, farmers are increasingly installing greenhouses controlled by IoT systems to monitor their crops in real time. However, using these data to generate a greenhouse climate model that allows intelligent and automatic control to reduce resources used while increasing crop production is challenging. Therefore, to develop this predictive model, the historical data set to train this model is crucial. These data are not available for the specific location where the greenhouse is installed until the IoT system starts operating. To solve the data problem, this study proposes the creation of synthetic greenhouse data using GAN techniques, to design a prediction system for climatic variables, specifically focusing on temperature, as it is one of the most influential monitored variables [21]. The findings of this study include:

- Creation of synthetic datasets using GANs techniques considering different time granularities.
- Study of the best prediction technique using neural networks to predict the temperature of a greenhouse, considering various granularities.
- Analysis and comparison of the different models created with both synthetic and original data, as well as with the fusion of both types of data.

The remainder of the paper is organized as follows. Section 2 summarizes state-of-the-art related studies regarding synthetic data generation in a time series. Section 3 describes the proposed GAN technique for creating synthetic time series data, as well as the techniques used for evaluating such synthetic data, including the description of the data and evaluation metrics used for the assessment. Section 4 shows the results, analysis and discussion. Section 5 highlights the conclusions and directions for future works

2 Related works

Data collection and capture is one of the mayor features of an open and well-served society. Innovative technologies allow us to capture, analyze and merge data from a variety of sources. However, data are not always accessible, because of privacy or because there is no local data collection system for a problem [22]. In this situation, new AI technologies provide tools and techniques capable of creating synthetic data. Synthetic data is a simulation of ground truth data that allows us to have a greater amount of information, to obtain more robust and accurate techniques [23]. When creating synthetic data, it is important to consider the type of data to be created. The creation of synthetic image data is useful and is widely used for health problems [24] or disease detection in crops [25]. However, the need for larger data sets is not exclusive to the world of image processing. Furthermore, in all contexts that require data for ad-hoc training, they also require large datasets, whether regarding IoT (where time series data predominates) or open contexts (where tabular data predominates). In [26], the authors review the role of IoT devices in smart greenhouses and precision agriculture, where variables such as the cost of agricultural production, environmental conservation, ecological degradation and sustainability have been analyzed. It shows how the economic benefits of using IoT applications in smart greenhouses have long-term benefits in commercial agriculture.

Focusing on the generation of synthetic data for time series data, synthetic data generation methods based on long-short term memory (LSTM) techniques are widely used. In [27], using LSTM, a method for completing synthetic well logs from existing log data was established. This method allowed, at no additional cost, synthetic logs to be generated from input log datasets, considering variation trend and context information. Furthermore, combining standard LSTM with a cascade system was proposed, demonstrating that this method gives better results than traditional neural network methods, and the cascade system improved the use of a stand-alone LSTM network, providing an accurate and cost-effective way to generate synthetic well logs.

Another of the most widely used techniques for synthetic data generation in recent years is GANs [28]. The use of GANs in time series has been widely used to detect anomalies, both in univariate [29–31] and multivariate models [32]. This scheme is widely used when working with unsupervised learning where anomaly detection is of particular importance for class labeling. The works on synthetic generation of time series data are not focused on agriculture; they are general works where techniques are proposed and evaluated with benchmarks or work focused on other areas. Yoon et al. [13] proposed a framework for the generation of synthetic time series data, where supervised and unsupervised techniques are combined. Specifically, the authors propose an unsupervised GAN with supervised training using autoregressive models.

However, in agriculture, using time series GANs is rarely used. Some studies have used agricultural data as benchmark data [33, 34], but to the best of our knowledge, there are no publications that focus on solving precision agriculture problems using GANs. In this study, the usefulness of synthetic data is investigated by assessing whether they preserve the distribution of individual attributes, the accuracy of the ML models and pairwise correlation.

3 Materials and methods

This section shows the datasets used and their characteristics. The synthetic data generation model was introduced before AI models were used to validate the effectiveness of the synthetic data described. Finally, different training strategies followed to achieve the objective are presented.

3.1 Dataset

The creation of synthetic data must first take a ground truth dataset from the particular domain for which synthetic data will be generated. In this case, the actual data are obtained from an operational greenhouse located in a semi-arid region of south-eastern Spain (Murcia). ground truth data is obtained from an IoT infrastructure that measures the inside temperature (ºC) of this greenhouse, which has been in continuous operation since 2018. This infrastructure sends 5 minutes of data grouped into 15-minutes, 30-minutes and 60-minutes respectively by performing the standard average.

Because the greenhouse is located in a semi-arid region, the thermal differences between summer and winter are remarkable; therefore, it has been considered that the ground truth data should be divided into winter and summer periods as well. Table 1 shows the ground truth datasets we have created for evaluation purposes. It shows the starting and ending date of the data, and the total number of values available. Datasets ending with a W indicate the end of the training data in winter and datasets ending with an S indicate the end of the training data in summer.

Table 1 Description of ground truth dataset

Datasets	Start date	End Date	# Instances
GreenHouse-15m-S	18-12-18	06-06-21	86543
GreenHouse-15m-W	18-12-18	17-01-21	73103
GreenHouse-30m-S	18-12-18	06-06-21	43272
GreenHouse-30m-W	18-12-18	17-01-21	36552
GreenHouse-60m-S	18-12-18	06-06-21	21636
GreenHouse-60m-W	18-12-18	17-01-21	18276

3.2 Synthetic data generation using GANs

For the generation of synthetic data, this study used *DoppelGANger*; a GAN architecture for sequential data proposed in [35]. Figure 1 shows the GAN architecture used that is based on the established architecture of strawman GANs for time series generation. It uses Recurrent Neural Networks (RNNs) to generate synthetic time series data. The generative part of *DoppelGANger* is based on a layer of LSTM cells with 100 units, following a batch generation strategy. Therefore, the model generates, in each pass, S consecutive records of the synthetic time series data, instead of a single one, as do most of the traditional approaches (e.g., R_1, R_2, .., R_S in Fig. 1). According to authors, this allows us to better capture the temporal correlation of long series and reduce the number of passes required by the model to generate the synthetic data. Furthermore, the GAN also includes a *normalization* mechanism for each input time series to tackle the well-known model-collapse problem of many GAN models. Then, the discriminator, which is a multilayer perceptron (MLP) with up to five layers of 200 neurons each followed by a ReLU activation function, uses the Wasserstein loss to report the differences between the ground truth and the fake data.

3.3 Deep Learning models

To assess the impact on the accuracy of ground truth and synthetic time series, four deep learning models have been considered: (1) MLP, (2) CNN, (3) LSTM and (4) a combination of CNN and LSTM.

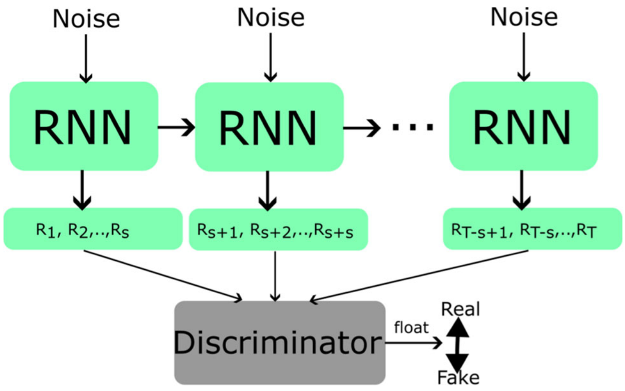


Fig. 1 Architecture of the DoppelGANger used for the synthetic data generation

- **MultiLayer Perceptron (MLP)**: The multilayer perceptron is an artificial neural network made up of multiple layers that forms a directed graph through the different connections between the neurons that make up the layers. This neural network attempts to simulate the biological behavior of neurons. MLP can solve non-linearly separable problems, because each neuron, apart from the inputs, has a non-linear activation function. The MLP is based on the backpropagation method. This method attempts to adjust the weights of the network connections to minimize the prediction error between the output produced by the network and the desired output. Layers can be classified into three types: The input layer comprises the neurons that input the data; no computation occurs in these neurons. Hidden layers can be as numerous as necessary depending on the complexity of the data; these layers comprise neurons whose input comes from previous layers and whose output and settings are passed on to subsequent layers. Finally, the output layer comprises neurons whose values correspond to the number of outputs of the network. In this study, a three-layer MLP comprising input, hidden and output layers are used. The first receives the input features; the hidden layer is where the inputs are processed so that the output layer generates the output of the MLP. The hidden layer learns any complex relationship between the input and the output due to the activation functions of its neurons [36].
- **Convolutional Neural Network (CNN)**: Convolutional neural networks are a type of supervised learning artificial neural network that processes its layers by mimicking the visual cortex of the human eye to identify different features in the inputs. These layers perform operations that modify the data to understand its particular characteristics. The three most common layers are: convolution, activation or ReLU, and clustering. The convolutional layer applies a set of convolutional filters to the input data where each filter activates different features. The rectified linear unit holds positive values and sets negative values to zero, allowing for faster and more efficient training, also known as activation, as only activated features proceed to the next layer. The clustering layer simplifies the output by a non-linear reduction of the sampling rate, which reduces the number of parameters the network must learn. These operations are repeated in tens or hundreds of layers; each layer learns to identify different features. After learning features in various layers, the architecture of a CNN moves on to classification. The penultimate layer is fully connected and generates a K-dimensional vector. The final layer of the CNN architecture uses a classification layer to provide the final classification output. The difference between a CNN and a traditional neural network is that a CNN has shared weights and bias values, which are the same for all hidden neurons in a given layer. Although the use of convolutional neural network models is more associated with the image classification domain, they are also used in different applications and domains, such as regression, where they can be used with time series by transforming the data to adapt them to the input of the convolutional network [37].
- **Long Short-Term Memory (LSTM)**: The LSTM model has a recurrent neural architecture with state memory, having the advantage of allowing long-term memory, and is therefore widely used in time series. LSTM is an evolution of standard recurrent neural networks, used in machine learning problems where time is involved, because their architecture as cells and loops allows the transmission and recall of information in different steps. LSTM comprises an architecture that allows information to be stored over long time intervals. This is because the memory cells of the network comprise several layers with loss functions (instead of one as in usual recurrent networks) of sigmoid type that allow us to bypass or add information to the main information line of the neural network, controlled by a hyperbolic tangent function. The information passes from one cell to another, first passing through a sigmoid layer, which is called the forget gate layer. It compares input and output, and returns a value between 0 and 1. If it is 1, the information is stored, if it is 0, it is disregarded. The next step comprises the second sigmoid layer and the hyperbolic tangent layer. It is used to decide which new information will be stored in the cell. The sigmoid layer called the input gate layer decides which value will be updated, and the hyperbolic tangent layer creates a vector of possible values decided by the previous one to be added to the state. The last step is a sigmoid layer that decides what the output will be, followed by a hyperbolic tangent layer that decides which values go to the network output according to the sign by which they are multiplied [38].
- **Convolutional Neural Network + Long Short-Term Memory (CNN+LSTM)**: This model, known as ConvLSTM, is a DL model that combines a CNN and an LSTM network. The architecture of this technique shares parts of the CNN and LSTM architectures with differences based on the connection point. In the CNN model, the fully connected end layer is replaced by the input layer of an LSTM. Thus the LSTM would keep its complete architecture, described above, and the CNN modifies its last layer. Therefore, the CNN network automatically extracts the input features, while the LSTM network obtains the regression results. This combination allows for the benefits of both models, creating a robust model for time series problems [39].

3.4 Preparation of datasets for training and testing

To accurately assess the impact of the synthetically generated data, five training and testing strategies are proposed to assess the performance of the ML models previously presented. The first strategy (that is, the *ground truth dataset*) is based only on the ground truth dataset (see Section 3.1). This dataset is divided into two datasets: (1) the training dataset, comprising all the data except the last day, and (2) the test dataset, comprising the last day of the available data. As these are time series data, it is impossible to perform a cross-validation or a validation with any other dataset than the latest values of the time series. time series require preserving the order and dependence between the data.

The second strategy for training and testing (namely, *Synthetic dataset*) only relays on the synthetic data generated with the GAN model previously presented. The synthetic dataset is divided into two datasets: (1) the data used for training, i.e., the synthetic data generated and the data used for testing that, in this case, are obtained from the ground truth dataset and (2) the data used for testing; i.e., the last day of the time series. The evaluation data are removed, and instead, the evaluation data are taken from the ground truth dataset, so the impact of the synthetic data on a real scenario can be rigorously evaluated.

The third strategy (namely, *Synthetic + Ground truth dataset*) combines synthetic and ground truth data. The ground truth dataset has been extended by adding data at the beginning of the dataset from the synthetic dataset to extend the time series and thus increasing the size of the dataset for training. Likewise, the models are trained using the entire dataset described above, removing the last day, which is reserved for testing.

The fourth strategy (namely, *Synthetic + Ground truth with reinforcement learning dataset*) is inspired by reinforcement learning. It also uses synthetic data with ground truth data but here, the training is performed by only using synthetic data. Once the model has been trained, the model is re-trained by using ground truth data. This is because the greenhouse will be continuously operating, and thus, data will be increasingly generated. Then, it can be used to increase the performance of the models over time. Likewise, the test strategy uses the last ground truth day to evaluate accuracy.

The fifth strategy (*Shuffled synthetic + Ground truth dataset*) uses synthetic and ground truth datasets. This test is like the third strategy, but, the synthetic dataset is shuffled before being concatenated at the beginning of the ground truth dataset. Like previous strategies, the last day of the ground truth dataset is used for testing. This strategy is used to verify the validity of a criterion-generated time series, and it would not be valid to introduce mere random data.

4 Evaluation and discussion

This study considers two dimensions of the problem: (1) the use of GANs for synthetic data generation (time series data) and (2) the impact on the accuracy of AI models depending on whether ground truth or synthetic data are used.

4.1 Exploratory data analisys

All the hyperparameters that have been used for using the GAN model are specified, described and explained in the following list:

- **Max sequence length**: Length of time series sequences, variable length sequences are not supported, so all training and generated data will have the same length sequences. Used value is: Lenght of the time serie for one day (96, 48 or 24), deppends on the dataset.
- **Sample length**: Time series steps to generate from each LSTM cell in DGAN, must be a divisor of max_sequence_len. Used value is: Lenght of the time serie for one day (96, 48 or 24), deppends on the dataset.
- **Batch size**: Number of examples used in batches, for both training and generation. Used value is: min(1000, length of the dataset).
- **Apply feature scaling**: Scale each continuous variable to [0,1] or [-1,1] (based on normalization param) before training and rescale to original range during generation. Used value is: True.
- **Apply example scaling**: Compute midpoint and halfrange (equivalent to min/max) for each time series variable and include these as additional attributes that are generated, this provides better support for time series with highly variable ranges. Used value is: False.
- **Use attribute discriminator**: Use separate discriminator only on attributes, helps DGAN match attribute distributions. Used value is: False.
- **Generator learning rate**: Learning rate for Adam optimizer. Used value is: 0.0001.
- **Discriminator learning rate**: Learning rate for Adam optimizer. Used value is: 0.0001.
- **Epochs**: Number of epochs to train model. Used value is: 100000.

Table 2 shows the main statistical values of the ground truth time series sampled every 15, 30 and 60 minutes during two and a half years together with the same descriptive statistics of the synthetic series over 288, 144 and 72 years.

Most are the usual statistical values. In particular, the standard error of the mean (SEM) measures how much discrepancy is likely in a sample's mean compared with the

Table 2 Comparison of ground truth and synthetic temperature time series distribution

	GreenHouse-15m-S		GreenHouse-30m-S		GreenHouse-60m-S		
	ground truth	Synthetic	ground truth	Synthetic	ground truth	Synthetic	RMSE
Samples	86,543	10,091,520	43,272	2,522,880	21,636	630,720	
Days	901,5	105,120	901.5	52.560	901.5	26.280	
Min	9.57	9.85	9.65	10.18	9.70	10.00	0.3847
Q1	17.68	18.05	17.70	17.46	17.72	17.33	0.3426
Median	20.99	21.44	21.00	20.83	20.98	20.85	0.2888
Mean	21.47	21.84	21.48	21.45	21.48	21.27	0.2459
Q3	24.91	25.34	24.89	25.26	24.89	24.74	0.3370
Max	33.72	33.68	33.71	33.65	33.70	33.61	0.0629
Sem	0.0170	0.0016	0.0241	0.0033	0.0339	0.0065	0.0218
Std	5.1992	4.9553	5.004	5.18	4.99	5.18	0.2051
Kurtosis	-0.52	-0.62	-0.52	-0.70	-0.52	-0.54	0.1194
Skew	0.34	0.27	0.34	0.32	0.34	0.31	0.0455
SMD (1)	0.073		0.006		0.041		

population mean. Kurtosis is the degree of peakedness of a distribution, if the value is close to 0, then a normal distribution is often assumed. Skewness is usually described as a measure of a dataset's symmetry, a value between -0.5 and 0.5, the data are fairly symmetrical. The statistics for skewness and kurtosis simply do not provide any useful information beyond that already given by the measures of location and dispersion but is another element to compare in the last column. Root-mean-square error (RMSE) is a frequently used measure of the differences between values, in our case ground truth and synthetic predicted values.

As can be observed, RMSE, calculated from the ground truth and synthetic column of each sampling rate, is a notably small value for all statistical measures shown. In addition, we can check the standardised mean difference (SMD) which tests for differences in means between ground truth and synthetic time series. Normally, a value of less than 0.1 is considered a "small" difference.

Table 2 shows a notably statistical similarity between the ground truth and synthetic values, especially because so many years are artificially generated. The data to see the distribution of the time series helps identify possible numerical anomalies such as outliers that would cause similar statistical values for different distributions. That is why these conclusions must be visually corroborated by looking at the box-and-whisker diagram shown in Fig. 2, the Kernel Density Function shown in Fig. 3 and the three Q-Q plots shown in Fig. 4 that compare the ground truth (line) and synthetic

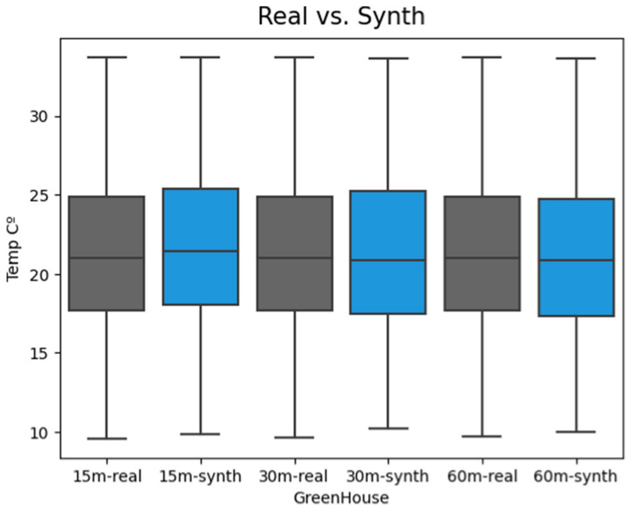


Fig. 2 Box plot comparing ground truth and synthetic data distributions according to sampling frequency

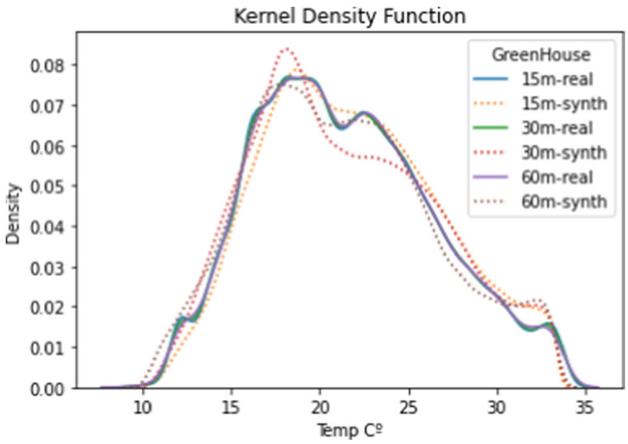


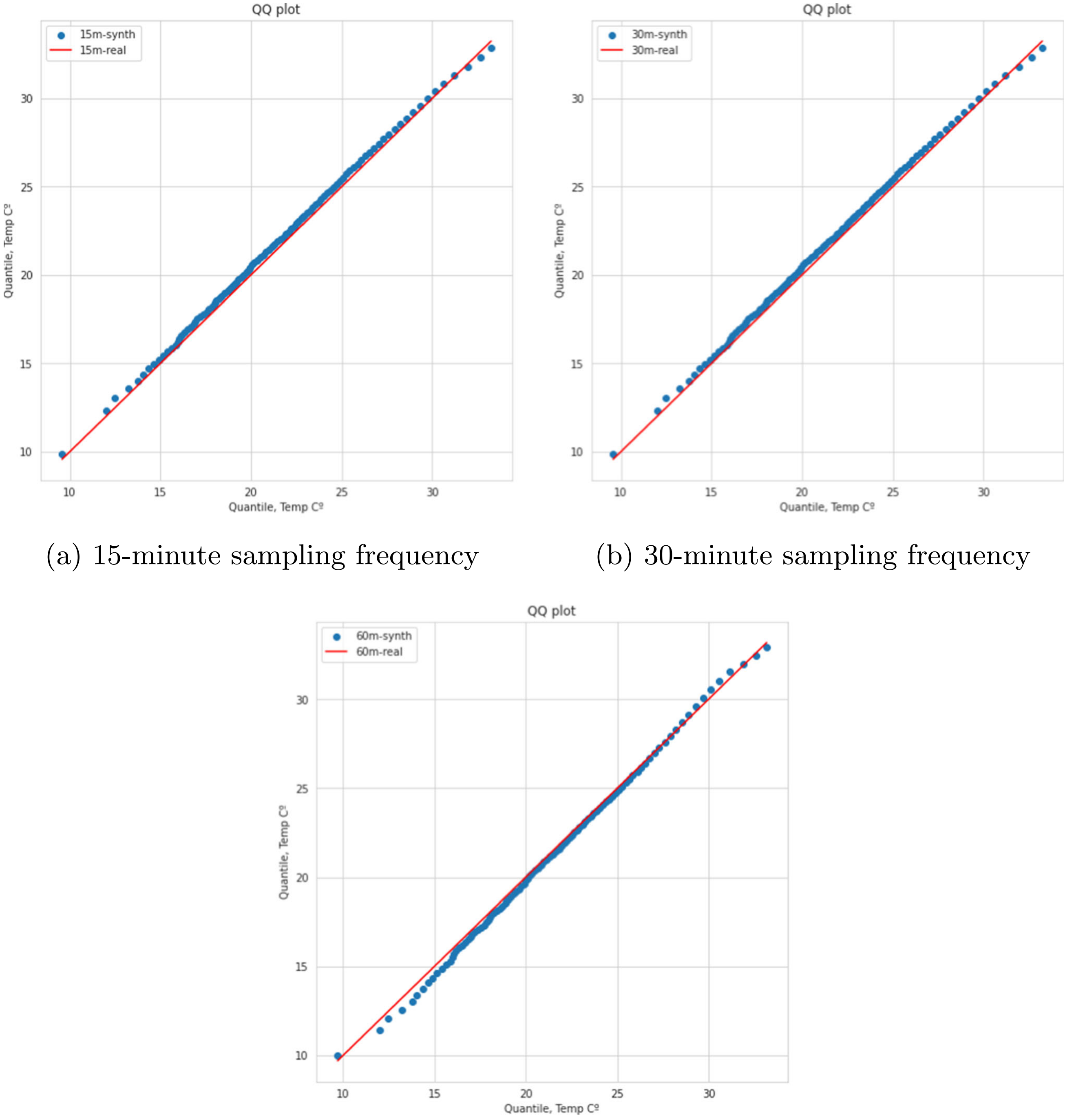
Fig. 3 Kernel density function comparing ground truth and synthetic data sets according to sampling frequency

(plots) sampling frequency for the three sampling rates.

$$SMD = \frac{\mid \bar{x_1} - \bar{x_2} \mid}{\sqrt{\frac{(s_1^2 + s_2^2)}{2}}} \tag{1}$$

To corroborate the conclusion that the generated synthetic time series will be useful to enrich the training of predictive models with tens of thousands of samples that we lack in reality, we compare on the timeline the three sets of generated series. Figure 5 shows a comparison of one week sampled every 15, 30 and 60 minutes between ground truth and synthetic data sets.

Visually, the synthetic time series is adjusted to the periodicity of each actual day. It is not perfect but significant



(a) 15-minute sampling frequency

(b) 30-minute sampling frequency

(c) 60-minute sampling frequency

Fig. 4 Q-Q plot comparing ground truth and synthetic data sets according to sampling frequency

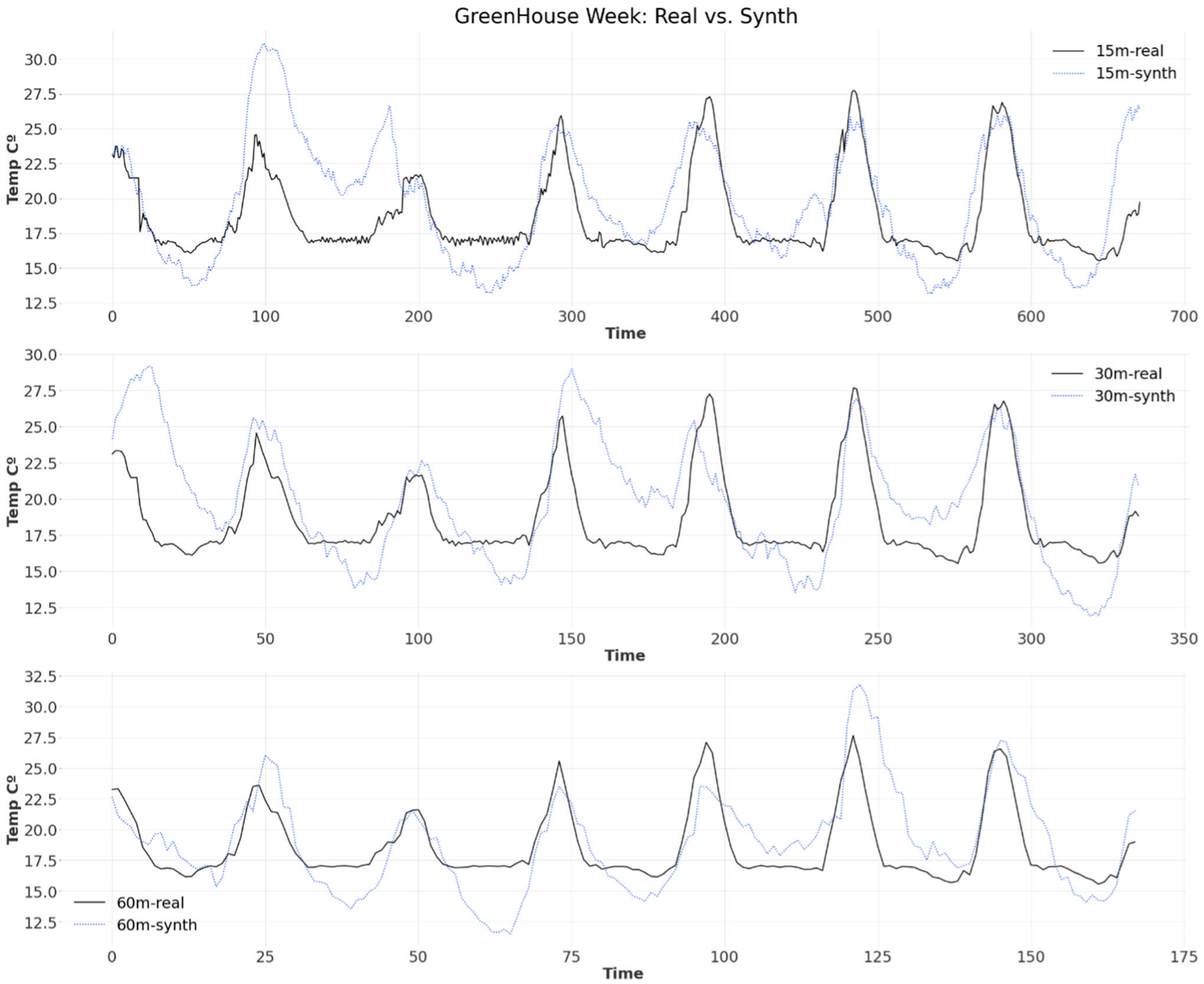


Fig. 5 Comparison of the same week of the three sampling rates with respect to their corresponding generated time series

correlations between each pair of ground truth and synthetic datasets are reported. However, they are not statistically significant when analyzing the correlation month-to-month or, year-to-year (see Table 3). A priori, this is not a problem for the intention to use the synthetic results to improve prediction models based on deep learning because the objective is to advance the prediction in a close time period.

In the following sections, this hypothesis is validated; i.e., that the generated data improve the training results of the proposed predictive model.

Table 3 Average of correlations between ground truth and synthetic data by time period and sampling frequency

Period of time	Data source	Corr. avg. GroundTruth-Synth
Day	GreenHouse-15m-S	0.8630
Day	GreenHouse-30m-S	0.8329
Day	GreenHouse-60m-S	0.7992
Month	GreenHouse-15m-S	0.5380
Month	GreenHouse-30m-S	0.5107
Month	GreenHouse-60m-S	0.4722

4.2 Model evaluation

Table 4 shows the models and hyperparameters used for assessment purposes.

The results of each model described in Section 3.3 using the above parameters are presented next. We have used three metrics to perform such an evaluation, the mean absolute error (MAE), the root mean squared error (RMSE) and coefficient of determination (R^2). These are some of the most common metrics used to measure accuracy for continuous variables. MAE and RMSE are suitable for model compar-

Table 4 Hyperparameters used for each model. (-) indicates model has no parameter

HYPERPAREMETER	DESCRIPTION	MLP	CNN	LSTM	CNNLSTM
Units	Number of neurons used in hidden layers	70	–	70	–
Filters	Features detector	–	64	–	64
Kernel size	Filters matrix used to extract the features from the dataset	–	1	–	2
Strides	Number of pixels shifts over the input matrix	–	1	–	4
Activation function	Function that decide if a neuron should be (or not) activated	Tanh	Tanh	Tanh	Tanh
Batch size	Size of bach used for training/forecasting	2880	2880	2880	2880
Epochs (+ *EarlyStopping*)	Number of epoch used in training	15000	15000	15000	15000
Optimizer	Function that optimises the learning of a artificial intelligence model, updating its neurons' weights depending on the error evaluation	Adam	Adam	Adam	Adam
Loss function	Function used for evaluate the error of the model in each epoch	MSE	MSE	MSE	MSE
Learning rate (+ *ReduceLROnPlateau*)	Percentage change with which weights are updated at each iteration	0.003	0.003	0.003	0.003

isons as they express the average model prediction error in units of the variable of interest. Their definition is as follows:

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

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

$$R^2 = \frac{\sum(e_i^2)}{\sum\limits_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

where, y_i is the real (ground truth) value of the climatological variable, $\hat{y}_i$ is the predicted value, e_i^2 is the error term and n is the number of observations.

Table 5 shows the values of the metrics for the MLP for the five train strategies described in seconds (secs.). 3.4. As seen, the strategy following a *reinforcement learning* approach achieved the best scores in most metrics and time horizons. This is especially remarkable for the datasets with a time frequency of 15 minutes (GreenHouse-15m-W and GreenHouse-15m-S). Furthermore, such a reinforcement approach provided more accurate MLP models than those solely relying on ground truth data. The R^2 of the former approach was 0.936 for GreenHouse-15m-S whereas the score of the latter strategy was only 0.644 given a 12-h time horizon. Similar behavior was observed for the 24-h period given the same dataset, 0.957 vs 0.835 R^2. The strategy using a shuffled version of the synthetic time series achieved larger errors than the one combining the time series because the GAN directly generated them. Concerning the sensitivity of the results, the accuracy of the MLPs trained following the *synthetic* or the *synthetic* + *ground truth* policies seem to slightly decrease with the frequency increases up to 60 min. For example, the R^2 score of the *synthetic dataset* MLP was 0.913 and 0.886 for frequencies 15 and 30 min given the summer dataset but it dropped to 0.749 when the frequency is set to 60 min. However, this pattern is not observed in the other policies in Table 5.

Table 6 shows the results obtained from the CNN model. Here, the three strategies that incorporated synthetic data during the training stage improve results than the one solely relying on the ground truth data. The combination of synthetic and ground truth data strategies achieved the best scores for all metrics and time horizons for the GreenHouse-15m-W feed. A similar behavior was observed in GreenHouse-30m-W. However, when the frequency increased to 60 min in the winter feed (GreenHouse-60m-W), *reinforcement learning* or only the use of synthetic data strategies provided better results. However, the summer datasets showed, a slightly different pattern. The CNN models trained with *synthetic* or the *reinforcement-learning* strategies were more accurate for the 30-min frequency (GreenHouse-30m-S dataset), but the combination of synthetic and ground truth strategy provided the most accurate CNN model for 15-min and 60-min frequencies. This reveals that combining the synthetic with the ground truth data approach improved the training of the

Table 5 Results of the MLP technique using ground truth, synthetic, a combination of ground truth + synthetic, ground truth + synthetic with reinforcement learning and shuffled synthetic + ground truth datasets

	Prediction hours	12h			24h		
	Metrics	R^2	RMSE	MAE	R^2	RMSE	MAE
ground truth dataset	GreenHouse-15m-W	0.787	1.543	1.260	0.865	1.643	1.256
	Greenhouse-15-S	0.644	3.755	3.445	0.835	3.379	3.001
	GreenHouse-30m-S	0.909	1.543	1.319	0.909	1.749	1.369
	GreenHouse-30m-S	0.479	3.939	3.518	0.808	3.205	2.717
	GreenHouse-60m-W	0.800	1.766	1.535	0.884	1.659	1.354
	GreenHouse-60m-S	0.831	3.448	3.048	0.822	3.512	3.157
Synthetic dataset	GreenHouse-15m-W	0.814	1.459	1.132	0.811	1.984	1.561
	GreenHouse-15m-S	0.913	1.863	1.535	0.951	1.497	1.164
	GreenHouse-30m-S	0.918	1.393	0.847	0.862	1.878	1.285
	GreenHouse-30m-S	0.886	2.890	2.362	0.941	3.761	3.188
	GreenHouse-60m-W	0.828	1.460	1.063	0.792	1.941	1.490
	GreenHouse-60m-S	0.749	2.913	2.558	0.878	2.764	2.438
Synthetic + ground truth dataset	GreenHouse-15m-W	0.818	1.899	1.487	0.890	1.647	1.316
	GreenHouse-15m-S	0.931	2.091	1.735	0.953	2.038	1.722
	GreenHouse-30m-W	0.858	1.276	1.004	0.908	1.317	1.066
	GreenHouse-30m-S	0.792	2.706	2.333	0.887	2.820	2.471
	GreenHouse-60m-W	0.741	1.849	1.499	0.788	2.094	1.767
	GreenHouse-60m-S	0.669	3.636	3.033	0.874	2.800	2.112
Synthetic + ground truth with reinforcement learning dataset	GreenHouse-15m-W	0.863	1.161	0.955	0.910	1.314	1.034
	GreenHouse-15m-S	0.936	1.678	1.358	0.957	1.330	1.025
	GreenHouse-30m-W	0.864	1.778	1.567	0.922	1.645	1.415
	GreenHouse-30m-S	0.816	2.685	2.350	0.921	2.629	2.306
	GreenHouse-60m-W	0.864	1.444	1.216	0.921	1.504	1.206
	GreenHouse-60m-S	0.898	2.708	2.074	0.938	3.257	2.782
Shuffled synthetic + ground truth dataset	GreenHouse-15m-W	0.912	1.430	1.100	0.899	1.750	1.283
	GreenHouse-15m-S	0.597	3.754	3.432	0.833	3.316	2.912
	GreenHouse-30m-W	0.814	1.727	1.482	0.900	1.660	1.325
	GreenHouse-30m-S	0.931	3.124	2.803	0.910	3.446	3.106
	GreenHouse-60m-W	0.841	1.673	1.488	0.898	1.656	1.331
	GreenHouse-60m-S	0.921	3.390	3.006	0.839	3.551	3.178

R^2 (coefficient of determination) RMSE (root mean square error) MAE (mean absolute error). RMSE and MAE are measured in degrees Celsius (°C) for each 12 and 24 hours. The best value for each combination of dataset, metric and prediction hour is shown in bold

CNN with high time frequencies (15 min) but for lower frequencies the other two synthetic-based approaches were also suitable. In terms of sensitivity, the models following *ground truth* or *shuffled synthetic+ground truth* approaches improve results when the frequency increases from 15 min to 60 min. However, the other three approaches follow the opposite trend with a slight accuracy improvement when decreasing the frequency of the time series (e.g. the R^2 score of the MLP with *Synthetic + ground truth* approach moved from 0.798 to 0.869 when the frequency of the GreenHouse-60m-S decreased from 60 to 30 min. This suggests that, for the MLP model, the combination of synthetic and real data must be better considered for time series with frequencies below 30 min.

Table 7 summarizes the evaluation of the LSTM model. The three synthetic-based training strategies outperformed the approach that only used ground truth data, considering most metrics, time horizons and datasets. For example, the RMSE of the LSTM trained only using ground truth data was 6.358 for the GreenHouse-15m-S dataset when considering a 24-h time horizon the same model trained with synthetic data achieved a much lower RMSE, 3.829.

Table 6 Results of the CNN technique using ground truth, synthetic, a combination of ground truth + synthetic, ground truth + synthetic with reinforcement learning and shuffled synthetic + ground truth datasets

	Prediction hours	12h			24h		
	Metrics	R^2	RMSE	MAE	R^2	RMSE	MAE
ground truth dataset	GreenHouse-15m-W	0.856	1.467	1.216	0.886	1.582	1.232
	GreenHouse-15m-S	0.516	3.809	3.434	0.808	3.199	2.715
	GreenHouse-30m-W	0.779	1.631	1.331	0.877	1.587	1.266
	GreenHouse-30m-S	0.922	3.313	3.056	0.914	3.368	3.027
	GreenHouse-60m-W	0.818	1.702	1.480	0.895	1.608	1.298
	GreenHouse-60m-S	0.860	3.396	3.082	0.859	3.430	3.091
Synthetic dataset	GreenHouse-15m-W	0.883	1.484	0.977	0.924	2.173	1.551
	GreenHouse-15m-S	0.906	3.334	2.123	0.714	4.890	4.063
	GreenHouse-30m-W	0.920	1.032	0.847	0.922	1.300	1.113
	GreenHouse-30m-S	0.928	2.382	1.758	0.938	2.521	2.155
	GreenHouse-60m-W	0.835	1.432	1.027	0.882	1.772	1.263
	GreenHouse-60m-S	0.716	3.057	2.683	0.892	2.789	2.435
Synthetic + ground truth dataset	GreenHouse-15m-W	0.897	1.121	0.823	0.915	1.194	0.908
	GreenHouse-15m-S	0.869	2.111	1.753	0.910	1.821	1.441
	GreenHouse-30m-W	0.930	0.946	0.722	0.932	1.036	0.843
	GreenHouse-30m-S	0.835	2.483	2.095	0.728	3.069	2.730
	GreenHouse-60m-W	0.808	1.659	1.099	0.882	2.411	1.741
	GreenHouse-60m-S	0.798	2.668	2.312	0.913	2.406	2.060
Synthetic + ground truth with reinforcement learning dataset	GreenHouse-15m-W	0.862	2.326	1.771	0.880	2.743	2.337
	GreenHouse-15m-S	0.951	2.251	1.833	0.942	2.297	2.014
	GreenHouse-30m-W	0.886	1.378	1.164	0.926	1.406	1.171
	GreenHouse-30m-S	0.911	2.555	2.127	0.854	2.038	1.557
	GreenHouse-60m-W	0.867	1.592	1.392	0.921	1.534	1.284
	GreenHouse-60m-S	0.795	2.737	2.375	0.924	2.714	2.353
Shuffled synthetic + ground truth dataset	GreenHouse-15m-W	0.779	1.852	1.605	0.859	1.692	1.413
	GreenHouse-15m-S	0.036	6.785	5.397	0.562	7.135	6.352
	GreenHouse-30m-W	0.793	1.899	1.648	0.877	1.691	1.397
	GreenHouse-30m-S	0.575	3.929	3.579	0.839	3.286	2.851
	GreenHouse-60m-W	0.890	1.486	1.249	0.913	1.654	1.273
	GreenHouse-60m-S	0.949	2.835	2.476	0.910	2.904	2.514

R^2 (coefficient of determination) RMSE (root mean square error) MAE (mean absolute error). RMSE and MAE are measured in degrees Celsius (°C) for each 12 and 24 hours. The best value for each combination of dataset, metric and prediction hour is shown in bold

Furthermore, the LSTM model exhibited differences in terms of accuracy depending on the time frequency of the model, as already observed with the CNN model. Therefore, Table 7 shows that the *reinforcement-learning* approach allowed the LSTM model to improve its accuracy for most of the datasets with low time frequencies (GreenHouse-30m-S, GreenHouse-60m-W and GreenHouse-60m-S). Furthermore, the approach that relies solely on synthetic data to train the model generated more accurate predictions datasets with higher time frequencies (i.e., GreenHouse-15m-W and GreenHouse-15m-S) at least for the 12-h time horizon. The training strategy based on a shuffled version of the synthetic time series achieved larger RMSE and MAE values than the three versions using the original synthetic time series, as well as the LSTM model just trained only with ground truth data. Table 7 also shows that all the models trained with the four policies, including ground truth data, were sensitive to the frequency of the input time series. The R^2 score exhibited an increase in the *ground truth*, *synthetic*, *synthetic + ground truth*, and *shuffled synthetic + ground truth* policies when the frequency of the time series moved from 30 to 60 min. In contrast, a different behavior was observed for the LSTM

Table 7 Results of the LSTM technique using ground truth, synthetic, a combination of ground truth + synthetic, ground truth + synthetic with reinforcement learning and shuffled synthetic + ground truth datasets

	Prediction hours	12h			24h		
	Metrics	R^2	RMSE	MAE	R^2	RMSE	MAE
ground truth dataset	GreenHouse-15m-W	0.321	2.723	2.055	0.420	3.857	2.887
	GreenHouse-15m-S	0.323	4.186	2.686	0.380	6.358	5.003
	GreenHouse-30m-W	0.826	1.478	1.176	0.890	1.618	1.288
	GreenHouse-30m-S	0.821	2.818	2.387	0.927	3.147	2.741
	GreenHouse-60m-W	0.815	1.746	1.520	0.901	1.637	1.293
	GreenHouse-60m-S	0.805	2.764	2.454	0.930	2.799	2.441
Synthetic dataset	GreenHouse-15m-W	0.890	1.456	0.995	0.908	2.364	1.771
	GreenHouse-15m-S	0.959	2.555	1.917	0.895	3.829	3.298
	GreenHouse-30m-W	0.982	1.236	0.948	0.801	1.853	1.528
	GreenHouse-30m-S	0.813	2.639	2.225	0.913	3.326	2.828
	GreenHouse-60m-W	0.979	1.832	1.307	0.873	2.011	1.476
	GreenHouse-60m-S	0.951	2.334	1.615	0.931	3.136	2.664
Synthetic + ground truth dataset	GreenHouse-15m-W	0.403	3.010	2.460	0.439	2.911	2.370
	GreenHouse-15m-S	0.225	4.552	3.866	0.123	4.642	4.021
	GreenHouse-30m-W	0.485	3.122	2.533	0.546	2.645	2.102
	GreenHouse-30m-S	0.862	2.865	1.913	0.893	2.291	1.645
	GreenHouse-60m-W	0.938	0.855	0.645	0.921	1.389	1.109
	GreenHouse-60m-S	0.931	2.418	1.596	0.945	2.478	2.035
Synthetic + ground truth with reinforcement learning dataset	GreenHouse-15m-W	0.863	1.475	1.251	0.923	1.435	1.163
	GreenHouse-15m-S	0.196	6.785	4.579	0.090	9.619	7.930
	GreenHouse-30m-W	0.801	1.840	1.483	0.860	1.631	1.280
	GreenHouse-30m-S	0.955	2.251	1.808	0.943	2.225	1.915
	GreenHouse-60m-W	0.879	1.316	1.061	0.934	1.285	0.979
	GreenHouse-60m-S	0.951	2.378	1.898	0.911	3.038	2.704
Shuffled synthetic + ground truth dataset	GreenHouse-15m-W	0.506	3.444	2.939	0.660	2.851	2.369
	GreenHouse-15m-S	0.570	3.281	2.751	0.651	3.748	3.253
	GreenHouse-30m-W	0.594	2.565	2.157	0.753	2.161	1.697
	GreenHouse-30m-S	0.726	3.083	2.693	0.900	2.786	2.401
	GreenHouse-60m-W	0.799	1.566	1.307	0.882	1.745	1.366
	GreenHouse-60m-S	0.857	2.814	2.521	0.942	3.038	2.690

R^2 (coefficient of determination) RMSE (root mean square error) MAE (mean absolute error). RMSE and MAE are measured in degrees Celsius (°C) for each 12 and 24 hours. The best value for each combination of dataset, metric and prediction hour is shown in bold

solely trained with synthetic data, its more accurate results were obtained with the frequency of the input time series was set to 15 min.

Last, Table 8 comprises the evaluation results of the CNN+LSTM model. The three training alternatives that used synthetic time series improved results, than the one that was based solely on ground truth data. Furthermore, we can see that the strategy that combined ground truth with synthetic data achieved the best results especially for the 15, or 30 min datasets. For example, the RMSE of the model for a 12-h prediction when trained was 0.932 for the GreenHouse-30m-W. This was a lower error than the one obtained by the variation trained only with ground truth data (i.e., 1.645). Furthermore, the CNN+LSTM model, trained only with synthetic data, achieves the best results for the two datasets with a 60-min frequency. Unlike the previous models, the *reinforcement-learning* strategy performed sligthly worse than the other alternatives. Moreover, the training using shuffled synthetic data, achieved slightly higher errors than the other four alternatives in most cases. Regarding sensitivity, CNN+LSTM variations improved scores with the 24-h time horizon than with the 12-h configuration. Furthermore,

Table 8 Results of the CNN+LSTM technique using ground truth, synthetic, a combination of ground truth + synthetic, ground truth + synthetic with reinforcement learning and shuffled synthetic + ground truth datasets

	Prediction hours	12h			24h		
	Metrics	R^2	RMSE	MAE	R^2	RMSE	MAE
ground truth dataset	GreenHouse-15m-W	0.881	1.565	1.364	0.918	1.548	1.245
	GreenHouse-15m-S	0.667	3.570	3.246	0.847	3.060	2.629
	GreenHouse-30m-W	0.862	1.645	1.448	0.916	1.619	1.300
	GreenHouse-30m-S	0.937	3.097	2.880	0.945	3.128	2.806
	GreenHouse-60m-W	0.882	1.411	1.244	0.928	1.493	1.209
	GreenHouse-60m-S	0.934	3.110	2.896	0.945	3.310	2.985
Synthetic dataset	GreenHouse-15m-W	0.746	2.603	2.015	0.753	2.019	1.495
	GreenHouse-15m-S	0.940	2.751	1.947	0.740	2.626	2.088
	GreenHouse-30m-W	0.885	1.042	0.852	0.935	1.052	0.825
	GreenHouse-30m-S	0.887	3.093	1.956	0.815	3.044	2.183
	GreenHouse-60m-W	0.863	1.156	0.860	0.906	1.260	1.008
	GreenHouse-60m-S	0.932	2.141	1.556	0.910	2.865	2.447
Synthetic + ground truth dataset	GreenHouse-15m-W	0.916	0.907	0.742	0.930	1.028	0.821
	GreenHouse-15m-S	0.935	2.416	2.175	0.926	1.972	1.664
	GreenHouse-30m-W	0.919	0.932	0.734	0.952	1.274	1.030
	GreenHouse-30m-S	0.938	1.667	1.239	0.896	1.594	1.207
	GreenHouse-60m-W	0.850	1.388	1.047	0.928	1.277	1.039
	GreenHouse-60m-S	0.886	2.017	1.695	0.853	2.413	2.146
Synthetic + ground truth with reinforcement learning dataset	GreenHouse-15m-W	0.851	1.498	1.269	0.924	1.398	1.117
	GreenHouse-15m-S	0.849	2.857	2.489	0.922	2.236	1.766
	GreenHouse-30m-W	0.831	1.583	1.339	0.914	1.593	1.277
	GreenHouse-30m-S	0.915	2.186	1.915	0.950	1.932	1.648
	GreenHouse-60m-W	0.843	1.382	1.152	0.919	1.435	1.131
	GreenHouse-60m-S	0.920	2.598	2.112	0.951	3.182	2.719
Shuffled synthetic + ground truth dataset	GreenHouse-15m-W	0.843	1.599	1.370	0.916	1.623	1.309
	GreenHouse-15m-S	0.921	3.249	3.043	0.948	3.132	2.823
	GreenHouse-30m-W	0.825	1.762	1.547	0.896	1.636	1.337
	GreenHouse-30m-S	0.918	3.354	3.097	0.900	3.344	3.018
	GreenHouse-60m-W	0.886	1.346	1.150	0.933	1.478	1.156
	GreenHouse-60m-S	0.885	3.326	2.888	0.853	3.578	3.183

R^2 (coefficient of determination) RMSE (root mean square error) MAE (mean absolute error). RMSE and MAE are measured in degrees Celsius (°C) for each 12 and 24 hours. The best value for each combination of dataset, metric and prediction hour is shown in bold

CNN+LSTM solely trained with ground truth data obtained better results for the summer than for the winter feeds considering its R^2 score (e.g., 0.928 vs 0.945 for the 60 min with 24 h as prediction horizon according to Table 8). This seasonal sensitivity was also observed in the other four policies incorporating synthetic data.

In this study, there are common patterns in the results of the four evaluated models. 1) The training of the forecasting algorithms leveraging the synthetic time series improved their prediction capabilities regarding the alternative of relying on ground truth data. 2) Common behavior is that using a shuffled version of the synthetic data did not provided no meaningful improvement regarding the models with just ground truth data. 3) The strategy combining ground truth with synthetic data provided the most robust models for 15-min and 30-min frequencies, at least for the CNN and LSTM variants. For larger frequencies, the *reinforcement learning* strategy provided more reliable predictors.

Evaluating the strategies has also revealed a sensitivity of the models to the frequency and season of the input time series. However, how these two factors affect the accuracy of the predictors strongly varies across models and training strategies with no global sensitivity pattern. Although the MLP and CNN with ground truth data performed better in the winter season, the other alternatives with synthetic data seem to provide better results in the summer time series. However, the CNN and CNN+LSTM alternatives do not follow such seasonal trends and show slightly better results in summer than in winter, regardless of the particular training strategy used to compose the predictor.

This has important implications in operational terms as it would be necessary to consider the relevance of the season and the frequency of the time series in order to eventually select a training strategy and the predictive algorithm. For example, in the case of greenhouse settings where the summer season was the most important part of the year, the evaluation showed that a CNN or CNN+LSTM instance trained with a *synthetic + ground truth* policy would be the most suitable configuration. The evaluation has shown that, for example, the RMSE of the CNN+LSTM model solely trained with ground truth data was above 3.00 for all the summer feeds (Table 8) whereas the CNN+LSTM fed with synthetic and ground truth data was below 2.42 for the same summer feeds.

These findings confirm the main hypothesis of this work, the usage of coherent synthetic time series, to enlarge the training sets of a forecasting model, helps to improve their final accuracy. Furthermore, shuffled series also shows that this improvement does not occur because we added more data to the training *corpus*, but because of the use of a synthetic series that actually behave in a similar manner to the target one.

5 Conclusion and future work

Precision agriculture is moving from tele-control systems to intelligent control systems by exploiting the data generated from the IoT system for a more sustainable and efficient crop management. This transition requires substantial amounts of reliable and ready-to-use data from the deployment of the system to train ML/DL models that meet expectations.

In this context, this novel study shows the reliability and suitability of using synthetic time series to expand the training *corpus* of deep-learning to forecast algorithms. The goal of these algorithms is to predict the internal temperature of greenhouses to anticipate future actions to keep this internal temperature within a suitable range. Five training strategies have been defined to optimally fuse ground truth and synthetic data.

The models trained with some of these fusion strategies outperformed the alternative models solely trained with the raw measurements from the temperature sensors by considering different time frequencies, evaluation metrics and time horizons. The metrics evaluated were affected by the frequency of the target time series and the season under consideration (winter or summer). This calls for a careful procedure to select the model and the training strategy based on the period of the year under study and the characteristics in terms of frequency and data curation applied on the input sequences of data.

This work opens a novel and promising research line for studying the most suitable training strategies for combining raw and synthetic time series in the development of a smart greenhouse. Future work will focus on: 1) Developing other combinations of ground truth and synthetic data to further improve the prediction of AI/ML models; 2) Using other synthetic data generation techniques and evaluating their effectiveness; 3) Apply the transfer learning technique for time series models of synthetic data generation; 4) Generate synthetic data and AI models in their multivariate version that consider all the variables that exist in a greenhouse; 5) Apply synthetic data generation methods and AI models in contexts other than those of precision agriculture in greenhouses.

Author Contributions Conceptualization, J.M.G., R.M.E, A.B.C. and F.T.S.; methodology, J.M.G. and F.T.S; software, J.M.G.; validation, J.M.C., R.M.E. and A.B.C.; formal analysis, A.B.C., J.M.C, and R.M.E; investigation, J.M.G., R.M.E, A.B.C., F.T.S. and F.A.T.; writing-original draft preparation, J.M.G., F.T.S., F.A.T., R.M.E, J.M.C and A.B.C.; writing-review and editing, A.B.C., R.M.E and J.M.C.; visualization, J.M.G and A.B.C.; supervision, J.M.C. and J.M.G.; project administration, J.M.C; funding acquisition, J.M.C. and A.B.C.

Funding This work is derived from R&D projects RTC2019-007159-5, as well as the Ramon y Cajal Grant RYC2018-025580-I, funded by MCIN/AEI/10.13039/501100011033, “FSE invest in your future” and “ERDF A way of making Europe” and the grant PID2020-112827GB-I00 funded by MCIN/AEI/10.13039/501100011033.

Data Availability All data and materials are available on request from the authors of this paper.

Declarations

Ethical Approval Not applicable

Consent to Publish All authors have read and agreed to the published version of the manuscript.

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Schepers J (2019) Precision agriculture for sustainability. Springer
2. Maraveas C (2022) Incorporating artificial intelligence technology in smart greenhouses: Current state of the art. Appl Sci 13(1):14
3. Wiebe K, Robinson S, Cattaneo A (2019) Climate change, agriculture and food security: impacts and the potential for adaptation and mitigation. Sustainable food and agriculture, 55–74
4. Guillén-Navarro MA, Martínez-España R, López B, Cecilia JM (2021) A high-performance iot solution to reduce frost damages in stone fruits. Concurr Comput Prac Experience 33(2):5299
5. Ardiansah I, Bafdal N, Suryadi E, Bono A (2020) Greenhouse monitoring and automation using arduino: a review on precision farming and internet of things (iot). Int J Adv Sci Eng Inf Technol 10(2):703–709
6. Castañeda-Miranda A, Castaño-Meneses VM (2020) Internet of things for smart farming and frost intelligent control in greenhouses. Comput Electron Agric 176:105614
7. Guillén-Navarro MA, Martínez-España R, Bueno-Crespo A, Morales- García J, Ayuso B, Cecilia JM (2020) A decision support system for water optimization in anti-frost techniques by sprinklers. Sensors 20(24):7129
8. Guillén-Navarro MA, Martínez-España R, Llanes A, Bueno-Crespo A, Cecilia JM (2020) A deep learning model to predict lower temperatures in agriculture. J Ambient Intel Smart Environ 12(1):21–34
9. Hewage P, Behera A, Trovati M, Pereira E, Ghahreman M, Palmieri F, Liu Y (2020) Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station. Soft Comput 24(21):16453–16482
10. Maraveas C, Asteris PG, Arvanitis KG, Bartzanas T, Loukatos D (2023) Application of bio and nature-inspired algorithms in agricultural engineering. Archives of Computational Methods in Engineering 30(3):1979–2012
11. Zhang C, Kuppannagari SR, Kannan R, Prasanna VK (2018) Generative adversarial network for synthetic time series data generation in smart grids. In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE
12. de Melo CM, Torralba A, Guibas L, DiCarlo J, Chellappa R, Hodgins J (2021) Next-generation deep learning based on simulators and synthetic data. Trends in cognitive sciences
13. Yoon J, Jarrett D, Van der Schaar M (2019) Time-series generative adversarial networks. Advances in neural information processing systems 32
14. Sorin V, Barash Y, Konen E, Klang E (2020) Creating artificial images for radiology applications using generative adversarial networks (gans)–a systematic review. Academic radiology 27(8):1175–1185
15. Chatziagapi A, Paraskevopoulos G, Sgouropoulos D, Pantazopoulos G, Nikandrou M, Giannakopoulos T, Katsamanis A, Potamianos A, Narayanan S (2019) Data augmentation using gans for speech emotion recognition. In: Interspeech, pp. 171–175 Springer Nature 2021 LATEX template Synthetic data generation for climate control in greenhouses 25
16. Pan Z, Yu W, Yi X, Khan A, Yuan F, Zheng Y (2019) Recent progress on generative adversarial networks (gans): A survey. IEEE Access 7:36322–36333
17. Lu Y, Chen D, Olaniyi E, Huang Y (2022) Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. Comput Electron Agric 200:107208
18. Wang C, Lee WS, Zou X, Choi D, Gan H, Diamond J (2018) Detection and counting of immature green citrus fruit based on the local binary patterns (lbp) feature using illumination-normalized images. Precis Agric 19(6):1062–1083
19. Lu Y, Young S (2020) A survey of public datasets for computer vision tasks in precision agriculture. Comput Electron Agric 178:105760
20. Choab N, Allouhi A, El Maakoul A, Kousksou T, Saadeddine S, Jamil A (2019) Review on greenhouse microclimate and application: Design parameters, thermal modeling and simulation, climate controlling technologies. Solar Energy 191:109–137
21. Cai W, Wei R, Xu L, Ding X (2022) A method for modelling greenhouse temperature using gradient boost decision tree. Information Processing In Agriculture 9(3):343–354
22. Choudhury SJ, Pal NR (2019) Imputation of missing data with neural networks for classification. Knowl-Based Syst 182: 104838
23. Nikolenko SI (2021) Synthetic Data for Deep Learning, vol 174. Springer, London
24. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F (2021) Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering 5(6):493–497
25. Abbas A, Jain S, Gour M, Vankudothu S (2021) Tomato plant disease detection using transfer learning with c-gan synthetic images. Comput Electron Agric 187:106279
26. Maraveas C, Piromalis D, Arvanitis K, Bartzanas T, Loukatos D (2022) Applications of iot for optimized greenhouse environment and resources management. Comput Electron Agric 198: 106993
27. Zhang D, Yuntian C, Jin M, Springer Nature 2021 LATEX template 26 Synthetic data generation for climate control in greenhouses (2018) Synthetic well logs generation via recurrent neural networks. Pet Explor Dev 45(4):629–639 (**Springer Nature 2021 LATEX template 26 Synthetic data generation for climate control in greenhouses**)
28. Figueira A, Vaz B (2022) Survey on synthetic data generation, evaluation methods and gans. Mathematics 10(15):2733
29. Lee C-K, Cheon Y-J, Hwang W-Y (2021) Studies on the gan-based anomaly detection methods for the time series data. IEEE Access 9:73201–73215
30. Jiang W, Hong Y, Zhou B, He X, Cheng C (2019) A gan-based anomaly detection approach for imbalanced industrial time series. IEEE Access 7:143608–143619
31. Niu Z, Yu K, Wu X (2020) Lstm-based vae-gan for time-series anomaly detection. Sensors 20(13):3738
32. Li D, Chen D, Jin B, Shi L, Goh J, Ng S-K (2019) Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: International Conference on Artificial Neural Networks, pp. 703–716. Springer
33. Han E, Crow WT, Holmes T, Bolten J (2014) Benchmarking a soil moisture data assimilation system for agricultural drought monitoring. J Hydrometeorol 15(3):1117–1134
34. Sanches I, Feitosa R, Achanccaray P, Montibeller B, Luiz A, Soares M, Prudente V, Vieira D, Maurano L (2018) Lem benchmark

database for tropical agricultural remote sensing application. International Archives of the Photogrammetry, Remote Sensing and Spatial

35. Lin Z, Jain A, Wang C, Fanti G, Sekar V (2020) Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In: Proceedings of the ACM Internet Measurement Conference. IMC –20, pp. 464–483. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3419394.3423643
36. Taud H, Mas J (2018) Multilayer perceptron (mlp). Geomatic approaches for modeling land change scenarios, 451–455
37. Li Z, Liu F, Yang W, Peng S, Zhou J (2021) A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems
38. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
39. Moishin M, Deo RC, Prasad R, Raj N, Abdulla S (2021) Designing deepbased learning flood forecast model with convlstm hybrid algorithm. IEEE Access 9:50982–50993

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Juan Morales-García is an Associate Professor and Researcher at the Computer Engineering Department, Catholic University of Murcia (Spain). He received his BSc, MSc and PhD degrees in Computer Science from the Catholic University of Murcia in 2020, 2021 and 2023, respectively. His current research interests include Artificial Intelligence (Machine Learning, Deep Learning), HPC, Edge Computing, IoT, and Pervasive computing.

Andrés Bueno-Crespo is an Associate Professor in the Department of Computer Engineering at the Catholic University of Murcia (Spain). He is a Computer Engineer and PhD in Computer Science (Polytechnic University of Cartagena, Spain, 2013). He is the author of more than 50 national and international publications in journals, book chapters, congresses and conferences, being oriented with the main lines of his research, mainly the use of Machine Learning and Deep Learning in different fields of applications. Currently, his research focuses on Deep Learning techniques for IoT infrastructures and pattern recognition in medical images. Dr. Bueno has worked on several competitive research projects and actively participates in the organisation of workshops and review processes of articles in international journals.

Fernando Terroso-Sáenz is an Associate Professor in the Department of Computer Engineering at the Catholic University of Murcia (Spain). In 2006, he graduated as a computer engineer. In 2005, he started its professional career acquiring experience in the management of geolocalized data. In 2013, he defended its doctoral thesis "Design of a context perception system for the intra-vehicular environment". In 2017, he joined the UCAM teaching in several subjects of the degree in computer science, as well as official masters.

Francisco Arcas-Túnez is an Associate Professor in the Department of Computer Engineering at the Catholic University of Murcia (Spain). His current research interests include the research in the field of natural language processing from the approach of deep semantics and machine learning, with a strong interest in its application to the needs of companies in this field and in any other where AI can be applied. Professor for more than 25 years in topics related to networking, cybersecurity and web and mobile application development.

Raquel Martínez-España is an Associate Professor in the Information and Communications Engineering department, faculty of Computer Science in University of Murcia, Spain. She obtained her M.S. in Computer Science in 2009 and her Ph.D. in Computer Science in 2014 at the University of Murcia. She has worked on several research projects in artificial intelligence and education. Raquel has participated in various academic and industry projects. Her research interests include machine learning, big data, soft computing, artificial intelligence, deep learning and intelligent data analysis.

José M. Cecilia is a Ramón y Cajal research fellow (Associate Professor Tenure track) at the Computer Engineering Department, Universitat Politècnica de València (Spain). He received his B.S. degree in Computer Science from Universidad de Murcia (Spain, 2005), M.S. degree in Computational Software Techniques in Engineering from Cranfield University (United Kingdom, 2007), and PhD degree in Computer Science from the Universidad de Murcia (Spain, 2011). Dr. Cecilia is a well-established researcher in the high-performance computing (HPC) area, specifically for the applications he has developed. His publication portfolio is composed of 50 research papers in JCR-Ranked journals, and 37 contributions on high-quality international conferences in the area (including ParCo, IJCNN, ICPP or GTC). Dr. Cecilia's research career has been focused on bridging the gaps between hardware and software by developing novel applications that overcome current technological barriers. Particularly, he has developed applications within the convergence between emergent HPC and Artificial Intelligence, following an application-driven approach where hardware and algorithmic innovations have been always considered to offer novel solutions to real problems.

Authors and Affiliations

Juan Morales-García[1] · **Andrés Bueno-Crespo**[1] · **Fernando Terroso-Sáenz**[1] · **Francisco Arcas-Túnez**[1] · **Raquel Martínez-España**[2] · **José M. Cecilia**[3]

Andrés Bueno-Crespo
abueno@ucam.edu

Fernando Terroso-Sáenz
fterroso@ucam.edu

Francisco Arcas-Túnez
farcas@ucam.edu

Raquel Martínez-España
raquel.m.e@um.es

José M. Cecilia
jmcecilia@disca.upv.es

[1] Computer Science Department, Catholic University of Murcia (UCAM), Av. de los Jerónimos, 135, 30107 Murcia, Murcia, Spain

[2] Information and Communications Engineering Department, University of Murcia (UM), C. Campus Universitario, 11, 30100 Murcia, Murcia, Spain

[3] Computer and Systems Informatics Department, Universitat Politècnica de València (UPV), Camino de Vera, s/n, 46022 Valencia, Valencia, Spain