Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hourly Solar Irradiation Forecast Using Hybrid Local Gravitational Clustering and Group Method of Data Handling Methods

Khalil BENMOUIZA (k.benmouiza@lagh-univ.dz)
Amar Telidji University of Laghouat https://orcid.org/0000-0001-5222-4055

Research Article

Keywords: Solar irradiation forecasting, local gravitational clustering, group method of data handling, clustering, phase space reconstruction.

Posted Date: December 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1050483/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Environmental Science and Pollution Research on April 15th, 2022. See the published version at https://doi.org/10.1007/s11356-022-20114-3.

Hourly Solar Irradiation Forecast Using Hybrid Local Gravitational Clustering and Group Method of Data Handling Methods

Khalil BENMOUIZA

Semiconductors and Functional Materials Laboratory, Amar Telidji University of Laghouat, Bp 37 C, Ghardaia Road, Laghouat 03000, Algeria

Tel: +213 778 73 85 56

Corresponding author

Khalil BENMOUIZA

E-mail address: k.benmouiza@lagh-univ.dz

Hourly Solar Irradiation Forecast Using Hybrid Local Gravitational Clustering and Group Method of Data Handling Methods.

Khalil BENMOUIZA

Semiconductors and Functional Materials Laboratory, Amar Telidji University of Laghouat, Bp 37 C, Ghardaia Road, Laghouat 03000, Algeria

Abstract

The foundation for many solar energies uses as well as economic and environmental concerns is global solar irradiation information. However, due to solar irradiation and measurements variations, reliable worldwide statistics on solar irradiation are frequently impossible or difficult to acquire. In addition, more precise forecast of solar irradiation plays an increasingly important role in electric energy planning and management due to integrating photovoltaic solar systems into power networks. Hence, this paper proposes a new hybrid model for 1-hour ahead solar irradiation forecasting called LGC- GMDH (local gravitational clustering- Group method of data handling). The novel LGC- GMDH model is based on the local clustering that adequately captures the underlying features of the solar irradiation time series. Each cluster is then forecasted using the GMDH method, which is a self-organized system that is capable of handling very complicated nonlinear problems. Finally, these local forecasts are reconstructed in order to obtain the global forecast. Comparative study between the proposed model and the traditional individual models such as; backpropagation neural network (BP), supporting vector machines (SVM), long short-term memory (LTSM), hybrid models such; BP-MLP, RNN-MLP, LSTM-MLP hybrid wavelet packet decomposition (WPD), convolutional neural network (CNN) with LSTM-MLP, and ANFIS clustering shows that the proposed model overcomes conventional model deficiencies and achieves more precise predicting outcome.

Keywords: Solar irradiation forecasting, local gravitational clustering, group method of data handling, clustering, phase space reconstruction.

1. Introduction

Nowadays, a natural desire to develop renewable energies exists. Indeed, we are talking about energy transition, which refers to the transition from the current energy system using non-renewable resources to an energy mix based mainly on renewable resources. Thus, the inclusion of renewable energies in the electricity grid is one of the current issues. Maintaining the balance of the network implies that at all times, electricity production is equal to electricity consumption. However, photovoltaic energy, like wind energy, is so-called intermittent energy. Indeed, it strongly depends on the weather. In order to be able to use renewable energies, solutions must therefore be found intermittently. The development of storage technologies, consumption control, and the development of the electricity network are thus envisaged. Likewise, the forecasting of photovoltaic production will have a primordial role in the future (Ahmed et al. 2020; Guermoui et al. 2020; Lai et al. 2021).

A critical step in predicting the amount of electricity produced by a solar panel is predicting the amount of energy received by the panel. Assuming that the inclination of the panels is low, we speak of a forecast of GHI (Global Horizontal Irradiance) which is the total irradiation received by a horizontal surface on the ground. The methodology used to predict irradiation depends on the desired forecast duration. Indeed, it is possible to use:

- Statistical models considering GHI measurements as time series. These are, for example, autoregressive moving average models (ARMA) (Voyant et al. 2012), autoregressive integrated moving average models (ARIMA) (Shadab et al. 2019), Kalman filter (Soubdhan et al. 2016), and Markov chain (Vindel and Polo 2014) are commonly used statistical techniques. These approaches are generally used for reasonably short forecast horizons. However, statistical models are unable to cope with non-stationary data that is abruptly changing. Solar irradiation data is non-stationary, resulting in poor forecast accuracy for sudden fluctuations.
- As more research is done, artificial intelligence (AI) and machine learning have gained much interest regarding solar irradiation forecast models. Artificial neural network (ANN) (Premalatha and Valan Arasu 2016; Benali et al. 2019) , support vector machines (SVM) (Wang et al. 2019) , extreme learning machine (ELM) (Burianek and Misak 2016), wavelet transform (Yadav and Behera 2014), deep learning (Lai et al. 2021), and ensemble learning are a few examples of typical AI techniques (Voyant et al. 2017; Alkhayat and Mehmood 2021). These methods can handle the non-linearity presented in the solar irradiation time series, and usually have high accuracy compared to statistical models.

- Numerical weather prediction (NWP) involves complex mathematical models used to simulate atmospheric change and forecast weather patterns using current weather circumstances. In meteorology, NWP models have met with a great deal of success, but because of their complexity, they cannot be easily implemented (Perez et al. 2013; Verbois et al. 2018).
- Physical models by observing clouds and their movement are used to forecast the amount of energy received on Earth (Chu et al. 2016; Blanc et al. 2017; Caldas and Alonso-Suárez 2019). This can be done by analysing images taken by ground cameras (forecast up to 30 minutes), or satellite imagery (forecast up to 5 hours).
- Combinations of these different models to take advantage of their complementarity create a category named hybrid forecasting models. Despite being very complicated, these models have become popular because they can incorporate a number of models and provide more accurate results for forecasting solar irradiation. Several models are proposed in different researches, a review of this category for forecasting solar irradiation is given in Refs. (Guermoui et al. 2020; Álvarez-Alvarado et al. 2021).

However, the solar irradiation is a dynamic time series representing high non-linearity (Gan et al. 2012), making it hard for the old forecasting models to obtain good results. Hence, as a solution clustering was introduced recently as a category of hybrid models for forecasting purpose (Benmouiza and Cheknane 2013, 2018) . Clustering (or data partitioning) is an unsupervised classification method that brings together a set of learning algorithms whose goal is to group unlabelled data with similar properties. In general, data are assigned to groups (clusters) so that observations within each group are similar to one another in terms of variables or attributes of interest, and the groups themselves are distinct from one another in terms of variables or attributes (Ghayekhloo et al. 2015; Benmouiza and Cheknane 2018).

Hence, based on this idea, we propose in this paper a hybrid model named LGC-GMDH in order to forecast the measured hourly solar irradiation data for one hour ahead. It consists of using a hybrid local gravitational clustering (LGC) (Wang et al. 2018) method and group method of data handling (GMDH) (Farlow 1981; Onwubolu 2016). The main idea consists of dividing the data set into training and testing set using the k-fold cross-validation technique (Kohavi 1995; Klipp et al. 2005). Then, the selected best set will be clustered into groups with the same proprieties using the LGC clustering. However, and due to the dynamic behaviour of the solar irradiation time series, the phase space reconstruction is used before the clustering phase (MacQueen 1967; Benmouiza and Cheknane 2013, 2018). It is a method used to reconstruct full system dynamics using a single time series. Takens (Takens 1981) explained that

104 a single vector of observations describing a chaotic system may be regenerated into a series of multidimensional
105 vectors using this theorem. As long as the embedding size is sufficient enough, the regenerated vectors may exhibit a
106 wide range of essential characteristics of the real-time series.

107 After that, each obtained cluster from the previous step will be forecasted as a single system using GMDH. It is
108 a multi-layered network with a predetermined structure. It offers the advantage of expressing nonlinear dynamics
109 mathematically and allowing higher-order terms without instabilities (Farlow 1981; Onwubolu 2016). GMDH is also
110 used to find input–output connections in models. It looks for connections between one output and a vast number of
111 potential inputs. The network determines which inputs are relevant to the specified system. During training, the
112 network is constructed layer by layer. Each layer contains neurons with just two inputs; their output is a quadratic
113 function of their two inputs.

114 Next, a global forecast is reconstructed based on each local forecast. Then, error metrics such as
115 Mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2) and forecasting skill
116 (FS) are used in order to compare the obtained forecast versus the measured test solar irradiation time series.

117 The main contributions of this paper are:

118 - A new multi-branch hybrid structure with clustered inputs is developed, which considers the highly
119 complicated dynamics of the hourly solar irradiation time series and exhibits outstanding prediction
120 performance.
121 - The used LGC algorithm in this paper introduces the data mining and clustering notions in order to forecast
122 locally the solar irradiation time clustered series, which has similar characteristics.
123 - The proposed LGC- GMDH framework enhance greatly the hourly-ahead solar irradiation forecasting taking
124 the advantages of feature selection using LGC algorithm and the power of the multilayer neural network
125 using GMDH with its predetermined structure.
126 - A comparison between the results obtained from our proposed LGC- GMDH model and several forecasting
127 models proposed in literature namely, single models such as; the Back Propagation (BP) neural network,
128 recurrent neural network (RNN), long short-term memory (LTSM), Support vector machines (SVM), and
129 hybrid models such as BP-MLP, RNN-MLP, LSTM-MLP and the hybrid wavelet packet decomposition

(WPD), convolutional neural network (CNN) with LSTM-MLP. Also, the ANFIS clustering model is used in this comparison in order to judge the goodness of our work.

The rest of this paper is organized as follows: Section 2 describes the principle of the proposed LGC- GMDH model. Sections 3 and 4 introduce the used solar irradiation data as well as the error metrics, respectively. Section 5 presents the simulation results and discussions. The comparison between the proposed model and the other models presented in the literature is shown in Section 6. The last section has devoted to the conclusion.

2. Methods

Our objective is to deliver the proper one-hour ahead forecast model in order to forecast the hourly solar irradiation time series. Hence, the LGC-GMDH model is proposed. This part will discuss the model framework as shown in Fig. 1, which summarises the adopted methodology of how this could work. The details of the whole process are given in the following steps:

- *Step A*: *data pre-processing*

The first stage for the proposed methodology is the data. It is necessary to collect and analyse the data before any prior use in the forecasting phase. The raw data should be divided into training and testing data sets. However, this data should first pre-process since it contains missing data due to the lack of measurement and suffers from the overfitting problem. To solve this, it was necessary to apply the k-fold cross-validation technique in order to solve the overfitting issue. Part (A) from Fig.1 depicts a conceptual model of cross-validation in its simplest form. First, the data were divided into k random subsets, S_1, S_2, and S_K, each of which was referred to as a fold. Each fold is about the same size as the previous one. The model was then run k times to see how well it worked. On each occasion, the k subset was utilized as a test set, while the other subsets were used as the training and validation sets, respectively. After that, the model with the most outstanding performance was selected from among the conducted k tests (Kohavi 1995; Refaeilzadeh et al. 2009; Wong 2015) .

- *Step B: data clustering*

The nonlinearity presented in the solar irradiation time series could affect the forecast model (Gan et al. 2012). In this step, the clustering was proposed as a solution to discover similar patterns from the solar irradiation time series. Clustering is an effective unsupervised learning method that has been used in various areas like pattern recognition,

image processing, bioinformatics, information retrieval, etc. Clustering is the process of separating a data collection using a similarity metric into uncoupled groups. The items are identical in the same cluster and the components are distinct in separate clusters (Benmouiza and Cheknane 2013, 2018). Several clustering techniques have been developed and categorized into categories such as partition-based clustering, hierarchical clustering, grid-based clustering, density-based clustering, model-based clustering, and so on (Wang et al. 2018).

However, before the clustering phase, the phase space reconstruction method is proposed in order to rebuild the global behaviour of a dynamic system from a single variable of the system in question (MacQueen 1967; Khalil and Ali 2016; Benmouiza and Cheknane 2018). Our goal is;

- Determine the minimum, appropriate, embedding dimension for phase space reconstruction for the time series (Kennel et al. 1992; Benmouiza and Cheknane 2018);
- Identify regions of the reconstructed phase-space which has similar characteristics using local gravitational clustering algorithm;

The theoretical background of both phase space reconstruction and LGC algorithm is given in detail in the following sections.

- *Step C: data forecasting*

In this step, the obtained clusters from the previous step will be forecasted using the GMDH (Farlow 1981; Onwubolu 2016). The main idea is to create local forecasts then a global forecast is reconstructed. The GMDH model is a self-organized system in which the structure optimizes itself based on the information provided by the user. This neural network attempts to build a function in a feed-forward neural network based on a second-degree transfer function using the GMDH neural network as the starting point. A computer program determines the number of hidden layers and neurons with the best deterministic transfer function in the generalized Markov decision-making network (GMDH). As a result, the best possible model structure is discovered. Several nonlinear functions, such as the Volterra series and the Kolmogorov-Gabor polynomial, are used to establish a connection between the input and output variables (Farlow 1981; Onwubolu 2016; Vaishnav and Vajpai 2018).

The error metrics will be used in order to judge the goodness of the forecast by comparing the obtained data with the measured one. Once the criteria are reached, the algorithm will end.

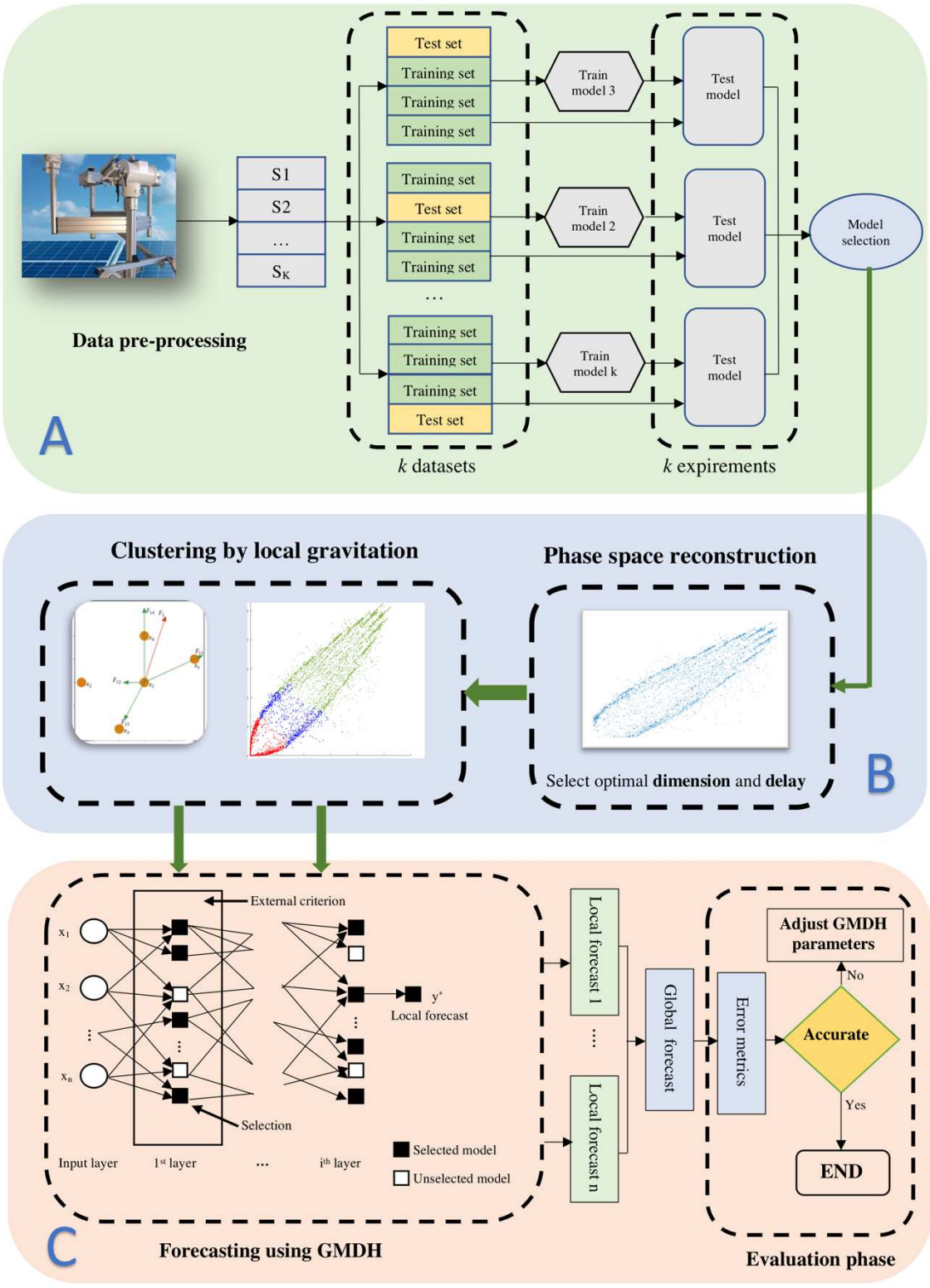

Fig.1 Flowchart of the proposed LGC- GMDH model.

In what follows, the mathematical modelling and the best parameters choosing for all the above-mentioned method are given in detail.

2.1 Phase space reconstruction

The phase space reconstruction introduced by Takens in 1981 (Takens 1981), is a technique that from a variable of a dynamic system, allows to reconstruct the global behaviour of its original system. It consists of embedding time series into high dimension space in order to understand its dynamics, which helps to provide a simple multidimensional representation of nonlinear time series. The principle of phase space reconstruction is based on the delay of the time series studied. We delay the time series in order to generate vectors of chosen dimensions which will constitute the reconstructed phase space (Benmouiza and Cheknane 2013, 2018).

The state vectors are defined as the set of d components corresponding to the d system's degrees of freedom. It is therefore theoretically necessary to have as many independent measurements as the system has degrees of freedom. In general, we will consider that we have only a sampled series resulting from the experimental observation of a fluctuating quantity and from which the system's phase space have to be estimated. A possible estimate consists in defining the set of vectors X_n ;

$$X_n = [x_n \ \dot{x_n} \ \ddot{x_n} \ \cdots]^T \tag{1}$$

Using finite differences, the downside of this method is its high sensitivity to noise. H. Whitney (Whitney 1936) proposed an alternative considering the vectors:

$$X_n = \left[x_n \ x_{n-\tau_1} \cdots x_{n-\tau_{d-1}} \right]^T \tag{2}$$

Where; d is the dimension of the estimated space and $\{\tau_i\}_{i=1,\dots,d-1}$ is a set of delays. d must verify $d \geq 2D + 1$, where D is the dimension of the state space (Whitney 1936). Authors in (Packard et al. 1980) have applied this method in the context of chaotic systems. A dynamic system will be said to be chaotic if it is a deterministic dynamic system whose behaviour is very dependent on the initial conditions in the sense that close initial trajectories diverge exponentially. The validity of the approach in the context of chaotic systems has been justified by F. Takens (Takens 1981) ; for an infinite series , for any values of τ_i and with the sufficient condition on the dimension d.

$$d \geq 2D_2 + 1 \tag{3}$$

Where D_2 is the correlation dimension (Grassberger and Procaccia 1983), the space formed by the set of vectors given in Eq.(2) is topologically and dynamically equivalent to the real phase space in the sense that the geometric properties of the attractor are preserved.

The phase space reconstruction technique involves two parameters to be adjusted. The first is the dimension of the phase space we want to generate, and the second is the delay we want to apply to the time series.

a) The choice of delay τ

In the case of non-noisy observations over infinite times, the choice of the delay τ is irrelevant (Takens 1981), but it becomes critical (Michel and Flandrin 1996; Benmouiza and Cheknane 2013, 2018) when the observation time is limited.

- If τ is too small, all the coordinates are strongly correlated: $x_k \approx x_{k+1}$, the vectors defined by Eq. (2) with $\tau_i = i\,\tau$ are almost collinear, the phase space tends to be a straight line, and the estimated dimension tends to be 1.
- If, on the contrary, τ is too large, the coordinates are almost independent. The set of vectors $\{X_n\}_{n=1,\dots,N}$ traverses almost all of the phase space: the system is close to a vector noise with independent coordinates, and the estimated dimension tends towards the value of the reconstruction dimension.

The most commonly used method is to choose τ at the value of the first zero of the estimated autocorrelation functions given by Eq. (4). It has also been proposed to take the first minimum of mutual information (Fraser and Swinney 1986) to determine τ.

$$C(\tau) = \frac{1}{N}\sum_{n=0}^{N} x_n x_{n+\tau} \tag{4}$$

b) Choice of dimension d

Let do $d_0 \in \mathbb{N}$ be the smallest dimension satisfying Eq. (3). If $d > d_0$, according to the theorem of F. Takens (Takens 1981), all the system dynamics are taken into account. Let $L = max_n x_n - min_n x_n$ and $N_{d\tau} = N - (d-1)\tau$, the density of observations per unit of hypervolume of dimension d is $\delta = \frac{N_{d\tau}}{L^d}$. Consequently, for a series of N observations, the probability density for an observation vector to be found in a given neighborhood of characteristic size l_0 typically varies as $\left(\frac{l_0}{L}\right)^d$. It follows that the estimation variance of the probability density on a partition of the

neighbourhood varying as the inverse of the number of points is expressed as a function of $\left(\frac{L}{l_0}\right)^d$ with the reconstruction dimension. It is therefore experimentally recommended to choose d as low as possible, i.e., $d = d_0$. Several methods have been proposed in the literature to determine this minimum value, among them;

- The estimation of the dimension of the attractor by the algorithm of P. Grassberger and I. Procaccia (Grassberger and Procaccia 1983) using the condition of Eq.(3),
- The study of the dynamics estimated by the reconstruction of different techniques such as singular value analysis (Auvergne and M. 1988), the estimation of the intrinsic local dimension (Michel and Flandrin 1996), the false nearest neighbours method (Kennel et al. 1992) and the vector field method (DT and L 1992). The dimension is then chosen as the smallest value d for which two points close in dimension d remain in dimension $d + 1$.

2.2 Clustering by local gravitation

Gravity theory has been used in clustering for a long time. Numerous gravity-based clustering algorithms have been developed, which replicate the process of objects being attracted and merging by their gravitational force. Each data point is considered an object by these methods and is assigned a mass in feature space. Local gravitation clustering LGC (Wang et al. 2018) is a method that uses local gravitational attraction to cluster data points, in which each data point is treated as a massed object that is attracted to its neighbours. According to (Wang et al. 2018), two key stages are included in the LGC Algorithm. First, LGC differentiated interior points, border points, and unlabelled points; and Second, LGC uses soft-connecting technologies to connect interior points. More details about this algorithm are given in what follows,

2.2.1 LGC Algorithm

Inspired by Newton's theory of gravity, the local gravitation in the data clustering technique reflects the relationship between a data point and its immediate neighbours. According to the theory of gravity, the attraction force between two-point masses may be calculated using the following formula:

$$\overrightarrow{F_{ij}} = G\frac{m_i m_j}{{D_{ij}}^2}\widehat{D_{ij}} \tag{5}$$

Where, $\overrightarrow{F_{ij}}$ is the force between two points i and j. D_{ij} is the distance between two mass points m_i and m_j. G is the gravitational constant and $\widehat{D_{ij}}$ is the line direction connecting two points along which the power acts. Because distances between points within the same local region are unlikely to change considerably, Eq. (5) can be simplified as follows:

$$\overrightarrow{F_{ij}} = Gm_i m_j \widehat{D_{ij}} \tag{6}$$

The resultant force of point i with its k-nearest neighbors (LRF) is :

$$\overrightarrow{LRF(i,k)} = \sum_{j=1}^{k} \overrightarrow{F_{ij}} = Gm_i \sum_{j=1}^{k} m_j \widehat{D_{ij}} \tag{7}$$

$\widehat{D_{ij}}$ contains the directional information between point i and its neighbours, and m_j values are considered as weighting factors in composing the forces in the neighbourhood

Eq. (7) shows that the neighbours of points with greater masses affect over those with smaller masses, and vice versa. Therefore, authors in (Wang et al. 2018) suggest that the new concept of the LRF replace Newton's theory of gravity as expressed in Eq. (8).

$$\overrightarrow{LRF(i,k)} = \frac{1}{m_i} \sum_{j=1}^{k} \widehat{D_{ij}} \tag{8}$$

Where the mas m_i is defined as ;

$$m_i = \frac{1}{\sum_{j=1}^{k} \widehat{D_{ij}}} \tag{9}$$

The Centrality (CE) based local clustering methods are proposed to take advantage of the LRF's information. Eq. (10) is the definition of the CE of data point i; k is the number of the neighbours.

$$CE_i = \sum_{j=1}^{k} cos\left(\overrightarrow{F_j}, \overrightarrow{D_{ij}}\right)/k \tag{10}$$

An LRF's approximate direction is indicated by a CE value larger than zero. Since

$$-1 \leq cos\left(\overrightarrow{F_j}, \overrightarrow{D_{ij}}\right) \leq 1 \tag{11}$$

And

$$-k \leq \sum_{j=1}^{k} cos\left(\overrightarrow{F_j}, \overrightarrow{D_{ij}}\right) \leq k \quad (12)$$

The property of CE is:

$$-1 \leq CE_i \leq 1 \quad (13)$$

According to Eq. (10), the neighbour's LRF direction is nearly opposite to D_{ij}, and CE is small if a point is at the border of a cluster. The opposite is true if a point is at the centre of a cluster.

In summary, data points in the centre area are differentiated from those in the border region using the physical principles based LGC technique, which calculates the LRF for each data point. In the first phase of the algorithm, LGC calculates the LRF field for every data point. After classifying the data points into the interior, border, and unmarked points, LGC performs the clustering process in the second stage based on LRF information and connects the internal data points. The data points that have been selected as limit points are used to form a cluster.

2.3 Group method of data handling (GMDH)

GMDH modelling technique is an effective method for the identification of higher-order nonlinear systems. A.G. Ivakhnenko was the first person to use this term (Farlow 1981). Furthermore, GMDH is an inductive self-organizing algebraic model, so it is not essential to know the precise physical model in advance to use it successfully. Instead, during the training phase, GMDH automatically discovers the dominant relationships among the system variables. In other words, the optimum neuron's structure is automatically chosen to minimize the values of the prediction error criterion, and any superfluous neurons are removed from the network. As a result, the GMDH has excellent generalization capabilities and can accommodate the complexity of nonlinear systems (Kondo 1998).

2.3.1 GMDH Model

The GMDH network is a data-driven modelling method. It uses mathematical functions to describe the complicated nonlinear connections between the provided input and output data sets. The GMDH network is comprised several layers, each of which contains neurons. Each neuron has two inputs and a single output (Nazerfard et al. 2006) . The output of each neuron is computed using the Ivakhnenko (Farlow 1981) polynomial stated in Eq. (14):

$$g(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 \quad (14)$$

The GMDH networks are constructed by combining the two input variables in each layer. The coefficients of the polynomials are then calculated for each combination and its associated output using the least square fitting method or regression analysis. To assess and verify the output of the polynomials once they have been computed, external accuracy requirements are used.

- **External criteria of accuracy**

The model adequacy is evaluated using the external criterion, which is known as the regularity criteria. It evaluates and measures the GMDH network's mean square error for each model neuron to assess the model output. When applied to the network, regularity also tells us which of the input combinations are more important. This method also evaluates the neuron polynomial's capacity and fitness to produce the intended system output. The closer the neuron polynomial fits the data, the lower the regularity requirement should be. The regularity criteria are given by Eq. (15).

$$R^2 = \frac{\sum_{i=1}^{N}(y_i - g_i)^2}{\sum_{i=1}^{N}(y_i)^2} \quad (15)$$

Where; R is the regularity criteria measure, N is the number of samples, y is the desired output, g is the GMDH neuron output.

Each neuron output is subjected to the regularity criteria in the GMDH network. In the next part, we will discuss the sorting process used to decide which neurons will survive to go on to the next layer.

- **Sorting Out Procedure Description**

Since the number of neurons and the number of layers is not specified, GMDH modeling is self-organizing. In order to ensure optimal performance, the best performing neurons in each layer of the GMDH are chosen using the external criterion defined in the preceding part. Neurons that have less than a pre-set number of periodicity criterion but which meet or exceed their established criteria on other neurons will be chosen for use. However, those that fall below the requirements set on other neurons will be removed and discarded from the network. Moreover, each of the layers keeps the lowest regularity criteria. When the following layer's lowest regularity criteria are greater than the preceding layer's smallest regularity criterion, adding layers will no longer produce new regularities. The last layer's neuron will provide the network's final output (Water et al. 2000). The main procedure for GMDH algorithm

implementation is described in the following steps (Nazerfard et al. 2006; Onwubolu 2016; Vaishnav and Vajpai 2018).

- *Step (1)*

The data set is separated into a training set and a checking set. The GMDH neuron's weights are estimated by using training data, while checking data is utilized to organize the network structures. A heuristic method is used to divide the data, either by choosing random locations for each group or by analysing the variance of the data (C. 2008).

- *Step (2)*

Generate every combination of two inputs conceivable among all of the input variables. The number of possible combinations is represented by $n = \frac{m(m-1)}{2}$, where m and n are the numbers of input variables and the number of combinations, respectively. After that, expand the inputs to a quadratic polynomial Z for each combination.

$$Z = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} & {x_{11}}^2 & {x_{21}}^2 \\ 1 & x_{12} & x_{22} & x_{12}x_{22} & {x_{12}}^2 & {x_{22}}^2 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{1N} & x_{2N} & x_{1N}x_{2N} & {x_{1N}}^2 & {x_{2N}}^2 \end{bmatrix} \quad (15)$$

In this way, the coefficients of the polynomial (g), which were previously stated in Eq.(14), are calculated for each combination in the training set. To get the polynomial weights, the least- squares fitting method is used, and the value obtained is given by Eq.(16).

$$W = (Z^T Z)^{-1} Z^T y \quad (16)$$

- *Step (3)*

The third step is to do a detailed evaluation and statistical testing of each polynomial using the testing data points. Eq. (17) is used to determine the outputs of each polynomial.

$$g = ZW \quad (17)$$

The criteria used to determine the regularity of each neuron in the first layer are determined after the polynomials have been computed. Neurons that meet the requirement for being "less than" a certain value are permitted to continue and become inputs for the neurons in the following layer. The remaining neurons are eliminated from the network.

- *Step (4)*

Finally, the whole process starting with the second step is repeated until the requirement for ending the GMDH network is fulfilled, at which point the procedure ends. When the lowest regularity criterion in the current layer is no longer smaller than the lowest regularity criterion in the preceding layer, the GMDH network will end. In order to get the final GMDH model, the route of the neurons in each layer that corresponds to the lowest regularity criterion is traced backward in time. The flowchart of the GMDH algorithm is given in Fig.2

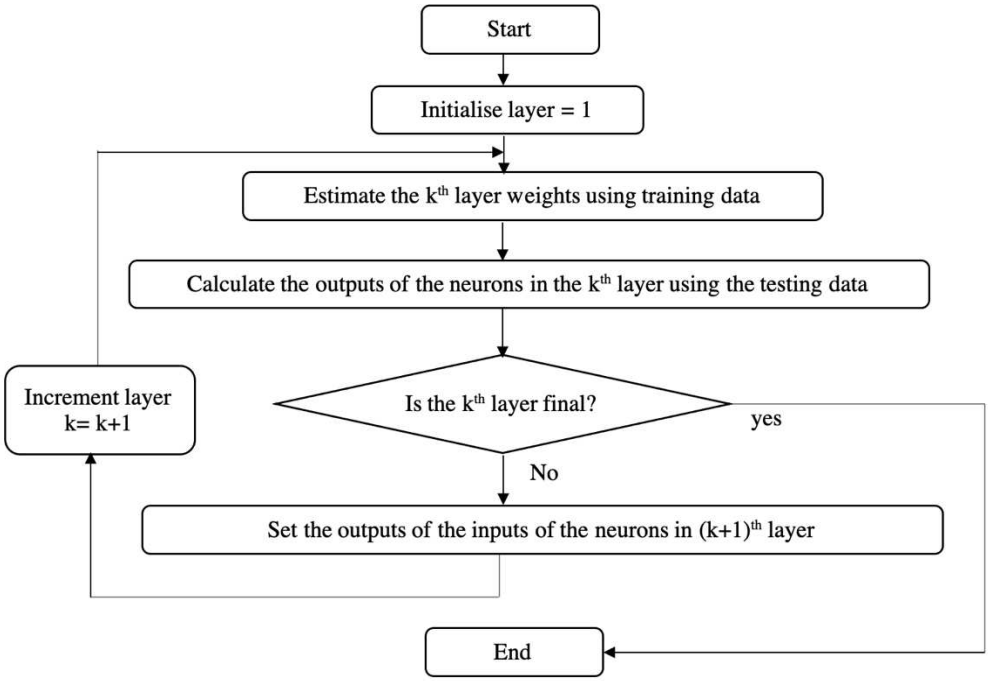

Fig.2 The GMDh network flowchart.

3. Data

The suggested approach was evaluated in Ghardaia, Algeria . It is situated in the heart of North Africa, along the Mediterranean coast, between the latitudes of 19° and 38° North and the longitudes of 8° and 12° East. A significant portion of the Sahara is located in its southern section (nearly 86 percent of the total area of the country). It is a key player in the area of solar technology due to its location in the sun belt and excellent climatic conditions. Algeria has the most significant solar energy potential in the MENA area. It has the highest effective solar potential where sunlight length is approximately 7.3 hours in the North, 8.3 hours in the highlands, and 10+ hours in the southern regions. As

a whole, the entire country's sunlight duration surpasses 3000 hours/year, while in the Sahara, it may approach 3500 hours/year (BENMOUIZA 2015).

This paper is especially interested in a particular area in the southwest of Algeria, namely the City of Ghardaia (positioned at 33.46° N, 3.78° E, at an altitude of 463 m). The accessible data is gathered by the enerMENA meteorological network project's meteostation (Schüler et al. 2016) . Ghardaïa meteorological station measures the global, DNI, and diffuse solar irradiation; furthermore, it monitors meteorological factors (air temperature, humidity, wind speed, atmospheric pressure). Hence, as a result, a database that's been around for over nine years enabled us to collect an in-depth knowledge of the area's solar potential. For DNI measurements, the meteostation is equipped with CHP1 pyrheliometers, and for diffuse (DHI) and global horizontal irradiation measurements, it is equipped with CMP11 thermal pyranometers (GHI). These instruments are placed on a Solys2 solar tracker with a sun sensor to provide continuous power. Collecting the data is accomplished via the use of a Campbell Scientific CR1000 data recorder.

4. Error metrics

A forecasting model's prediction values must be tested to determine its performance. To find the optimal solution, calculate appropriate error metrics. A model's error metric is a method for measuring the quality of a model and may be used by the forecaster to compare several models. It allows us to assess better how effectively the model completes its various duties. To this end, we have chosen four metrics in order to evaluate the forecasted data noted as $\hat{y}$ versus measured one noted y and a number of observations noted N. These statical metrics are summarized as follows (Benmouiza and Cheknane 2018; Botchkarev 2019; Huang et al. 2021) ;

4.1 RMSE

In mathematics, root means square error (RMSE) is defined as the square of the average of the errors' squares. It is also described as a measure used to evaluate the quality of a forecasting model or predictor, among other things. RMSE takes into account both variance (the dispersion of expected values from one another) and bias (the difference between two anticipated values) (the distance of predicted value from its true value). This error measure is likewise always positive, with lower values being preferable. The RMSE number is in the same unit as the predicted value, which is a benefit of this approach. In comparison to Mean square error MSE, this makes it simpler to comprehend. The RMSE may also be compared to the Mean Absolute Error MAE to see whether the prediction includes significant

but occasional mistakes. The greater the gap between RMSE and MAE, the greater the inconsistency of error magnitude. This measure may gloss over issues related to poor data volume. Eq. (18) gives the calculation of RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{N}} \tag{18}$$

4.2 Mean Absolute Error MAE

MAE is the average of the absolute difference between forecasted and actual values. The MAE informs us how much inaccuracy we may anticipate from the prediction on average. The error numbers are in the predicted values original units, and MAE = 0 implies that there is no mistake in the forecasted values. The lower the MAE number, the better the model; a value of 0 indicates that the prediction is error-free. In other words, when comparing several models, the model with the lowest MAE is seen to be superior. However, MAE does not show the proportional magnitude of the mistake, making it impossible to distinguish between large and minor errors. It may be used with other measures (Root Mean Square Error) to assess whether the mistakes are greater. Furthermore, MAE may obscure issues related to low data volume. The MAE is calculated using the Eq.(19);

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \tag{19}$$

4.3 R-squared value

R-squared or R^2 is a statistical measurement that illustrates how near the data are to the fitted regression line. It is sometimes known as the coefficient of determination, or the coefficient of multiple determination for multiple regression or the correlation coefficient. R-squared represents the proportion of the response variable variance described by a linear model. R^2 values are between 0 and 100%. A correlation coefficient between 0% and 100% implies that the model has little explanatory power in explaining the variability of the response data around their mean. The 0% indicated that there is no fluctuation in the response data around its mean; therefore, the model has explained nothing. When 100% appears, the model explains all of the response data's variability around its mean. Hence higher R^2 values, the better the model fits the data. R^2 is calculated using Eq. (20);

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}\right) \tag{20}$$

4.4 Forecasting skill

The forecasting skill (FS) is defined as the degree of accuracy with which a prediction is associated with estimating of actual values. It is calculated using Eq.(21) (Mazorra Aguiar et al. 2015; Schmidt et al. 2016) ;

$$FS = 1 - \left(\frac{RMSE}{RMSE_{smart}}\right) \tag{21}$$

$RMSE_{smart}$ is the smart persistence model, which consists of forecasting clear sky index for each time horizon h persists for the next time step given by;

$$k_t^*(t+h) = k_t^*(t) GHI_{clear}(t+\Delta t) \tag{22}$$

k_t^* is the clear sky index given by;

$$k_t^* = \frac{GHI}{GHI_{clear}} \tag{23}$$

Where ; GHI is the measured horizontal hourly solar radiation data at ground level and GHIclear is the calculated clear sky hourly solar radiation data (W.M.O. 1981; Tadj et al. 2014).

A perfect forecast will have an FS equals 1, and an FS equals 0 means there is no improvement over the reference model.

5 Simulation results

The main objective is to provide the best model for 1-hour ahead GHI forecasting. Hence, we have proposed the clustering using by local gravitation method and the forecasting using GMDH. To this end, we will first train the model from 1st January 2020 to 30th November 2020, and test it for 744 hours (total predicted hours) from 1st December 2020 to 31st December 2020. Then, two years have been chosen for training and 6 months for testing.

First, and using the time delay embedding technique, the phase space reconstruction of hourly solar irradiation was obtained. We determined the best time delay for the phase space reconstruction to be one, with dimensions equal to two. Next, the LGC algorithm was applied to the results obtained from the phase space reconstruction. The clustering results from the LGC show that the number of clusters is 3.

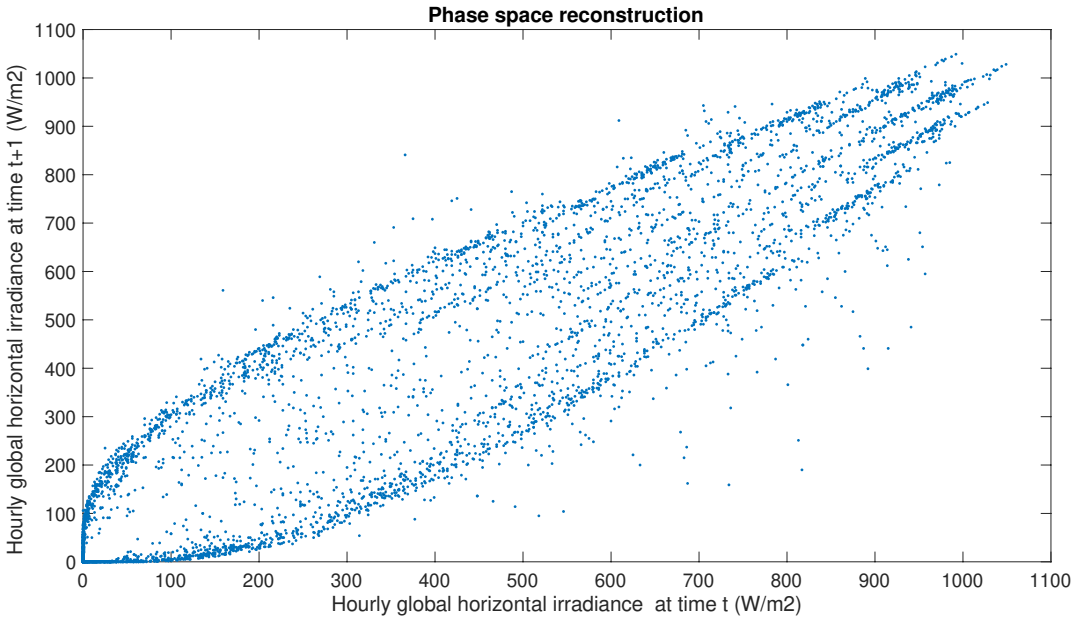

Fig.3 Phase space reconstruction of the GHI at time t and t+1 (1st January to 30th November 2020, Ghardaia).

The simulation results of the reconstructed phase space, the LGC clustering, and the three subsets obtained from the LGC clustering algorithm of the solar irradiation are shown in Figs. 3 ,4, and 5, respectively.

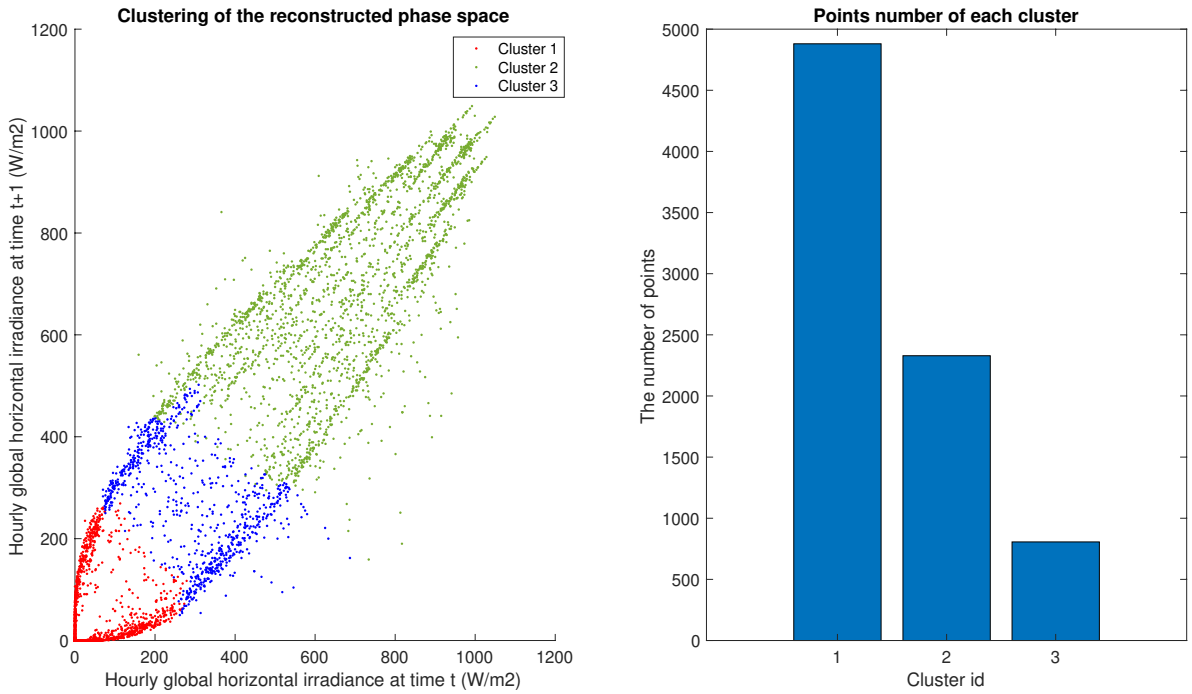

Fig.4 LGC for the reconstructed phase space of the training data (1st January to 30th November 2020, Ghardaia).

These figures show that the solar irradiation time series had been grouped into three clusters. High values solar irradiation (cluster N°2) that describe the number of sunny hours throughout the year (for example, the number of hours in the middle of the day when there are no clouds) are contained in this cluster. These values are contrasted with medium values solar irradiation (cluster N°3) that depict hours when the sky is partly cloudy (for example, hours from 9 AM to 11 AM and from 2 PM to 4 PM). Finally, values of low solar irradiation (cluster N°1) describe hours when the sky is entirely cloudy (for example, when it is raining or snowing) (or the hours of sunshine and sunset).

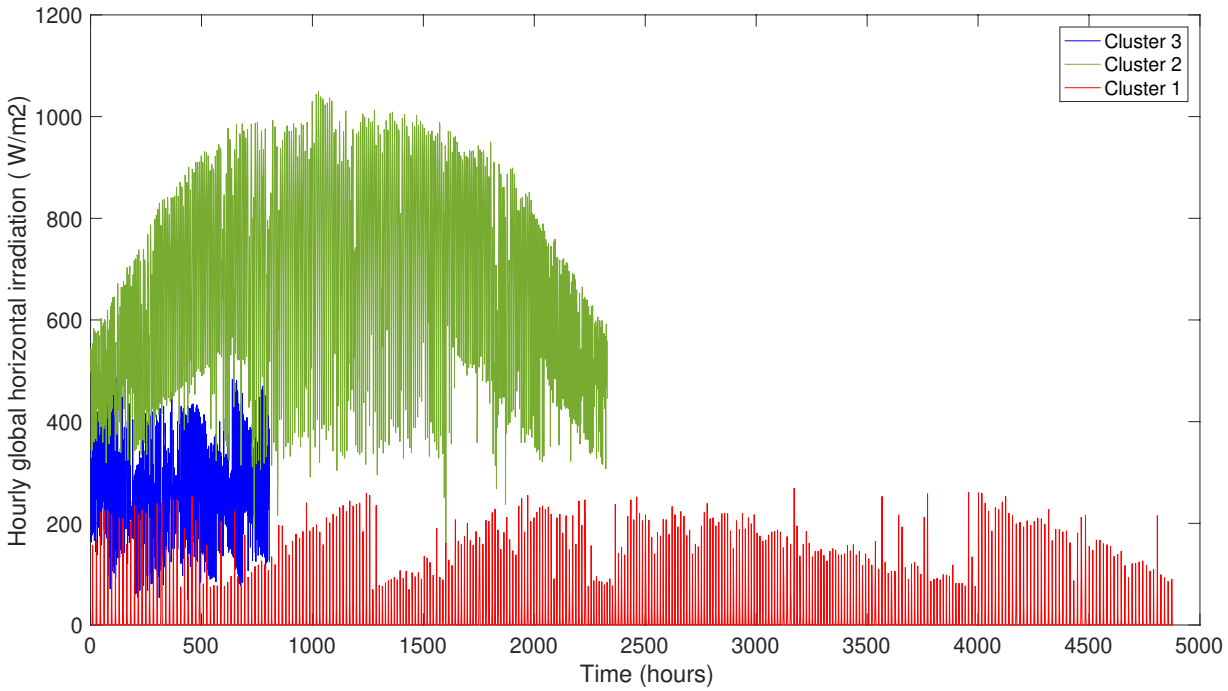

Fig.5 Obtained clusters from the LGC.

The next step consists of forecasting each cluster alone, then, obtain the global forecast. In the simulation, various numbers of parameters are chosen in order to achieve the best forecast. The calculation of errors has been used to assess the performance of the system. The simulation results for the best configuration of the forecasted and measured test data are shown in Fig.6 for December 2020; the red line represents the forecasted data, and the blue dot line shows the measured hourly global solar irradiation data. Furthermore, the relative error and its histogram are also shown in this figure.

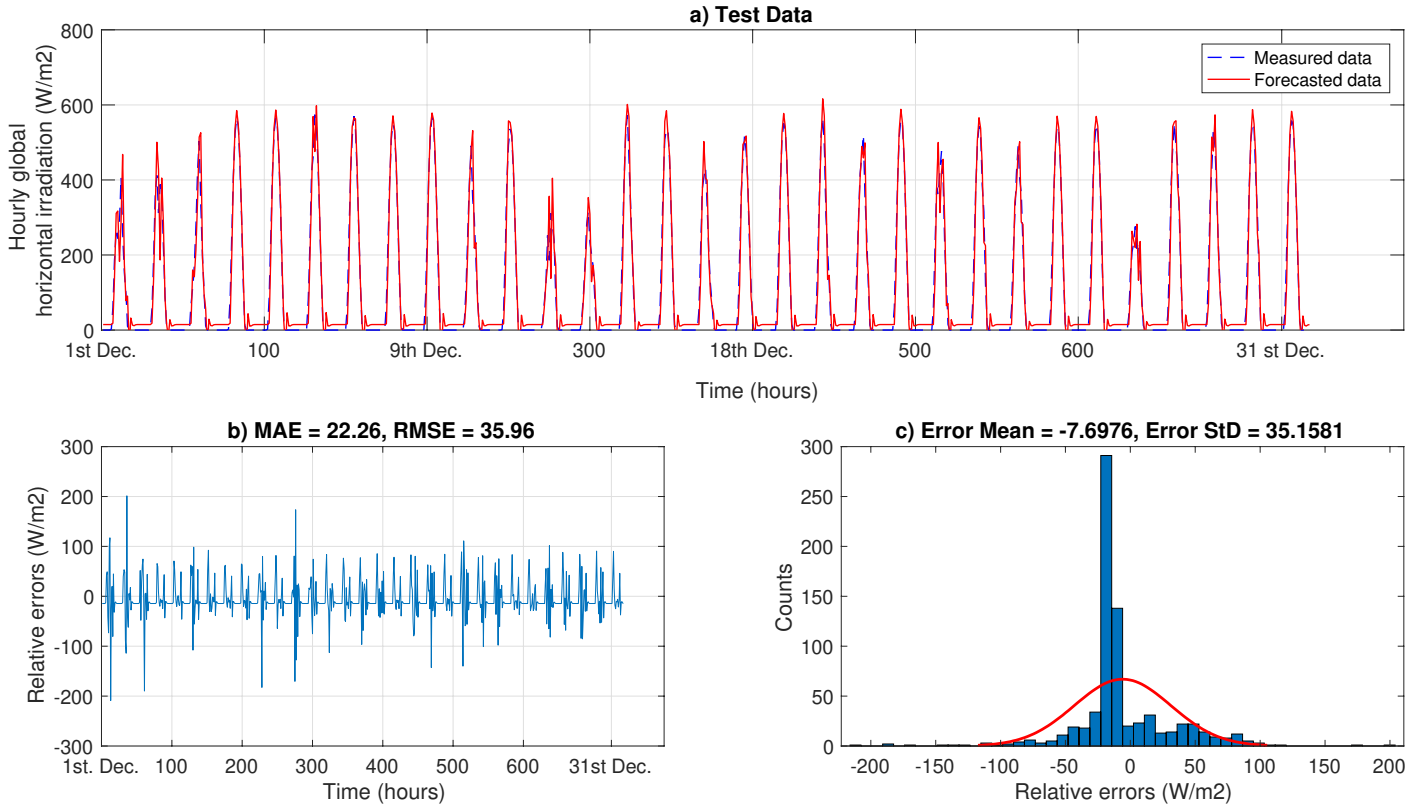

Fig.6 a) Forecasted and measured testing data for December 2020 using hybrid LGC with GMDH model, b) the relative error, and c) the histogram.

In addition, the training and testing data versus its fit with their R-squared values are shown in Fig.7. The majority of the locations in both the predicted and measured series are within a few degrees of each other. Because of the large number of cloudy days, there are some delays in the system's response time.

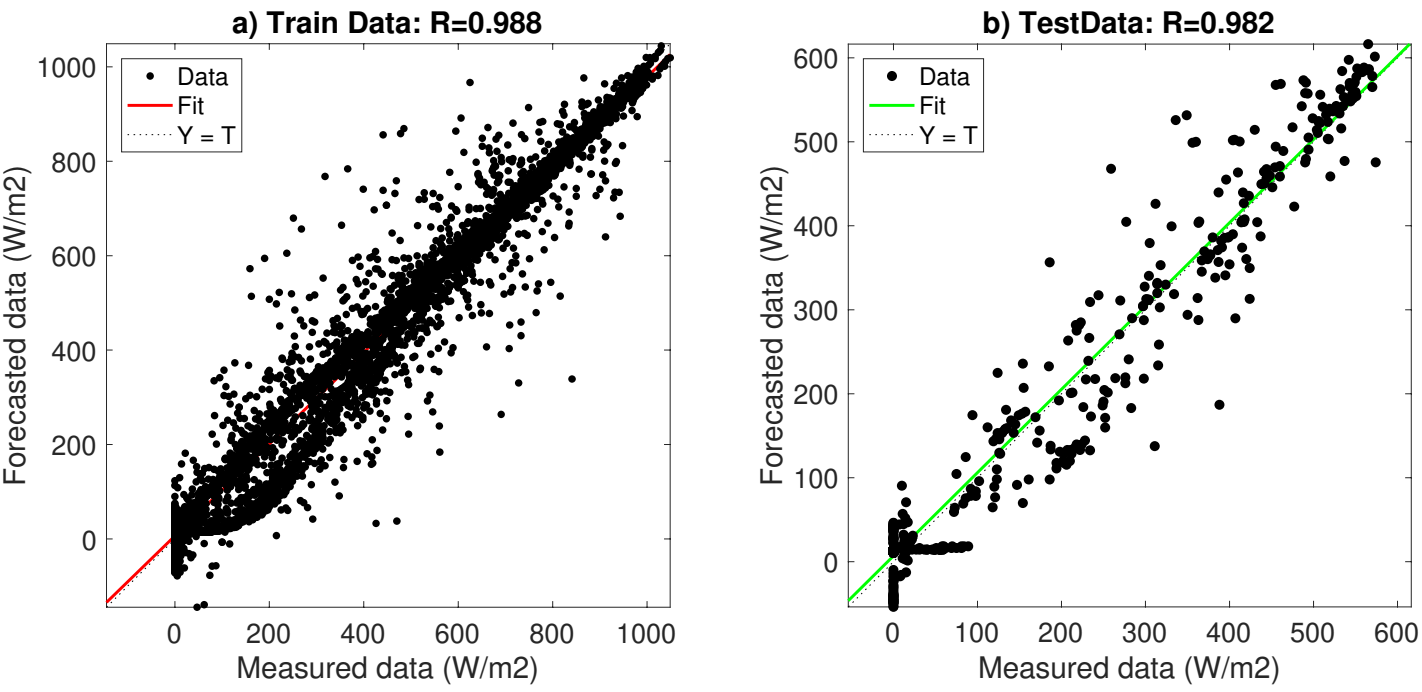

Fig.7 The measured data versus estimated training (a) and testing (b) for Ghardaia 2020 using hybrid LGC with GMDH model.

A further evaluation of the forecasted hourly global horizontal time series was carried out by computing the root mean squared errors (RMSE), mean absolute error (MAE), the R^2 value, and forecasting skill for training and testing data set between the actual data and the forecasted data for the period from the first of December 2020 to the last day of December 2020. Several tests were performed by changing the number of neurons and layers for each forecast. The results are expressed in Table 1.

Table 1 RMSE, MAE, R^2, and FS results between measured and forecasted data for different neurones and layers.

layers	neurons	Training data			Testing data			
		RMSE (W/m^2)	MAE (W/m^2)	R^2	RMSE (W/m^2)	MAE (W/m^2)	R^2	FS (%)
5	25	51.61	31.46	0.986	37.59	26.43	0.978	47.85
3	10	53.46	33.96	0.985	37.90	27.48	0.978	47.79
20	50	49.65	29.12	0.987	36.71	25.18	0.980	47.98
60	100	48.61	27.26	0.988	35.96	22.26	0.982	48.12

Following the results of Table 1, we can see clearly that a high number of layers and neurons gives the lowest errors and highest R^2 and forecasting skills. However, it takes more computational time. Hence, it is not necessary to increase their number since the results are close to each other. Moreover, from the results of Table 1 and Fig.6 which represent the plots of the best configuration we can see an overall RMSE was equal 35.96 W/m^2 and the MAE equals 22.26 W/m^2, which may be regarded as excellent predicted values.

Furthermore, based on Fig. 7, the R-squared value computed by Eq. (20) is equal 0.982, which is a positive value with a good forecasting skill of 48.12%. Finally, based on the simulation findings, it was determined that this approach would be an excellent way to forecast solar radiation data.

In the same way, we have tried larger data sets in order to forecast more than one month. Hence, we have chosen a solar irradiation time series measurement for three years from the 1st January 2018 to the 31st December 2020 for the site of Ghardaia, Algeria. Thirty-three months were used for training and 3 months were used as forecast horizon. The same methodology applied for 1 month forecast has been applied for the case of three months. The results of the clustering of the reconstructed phase space are shown in Fig. 8.

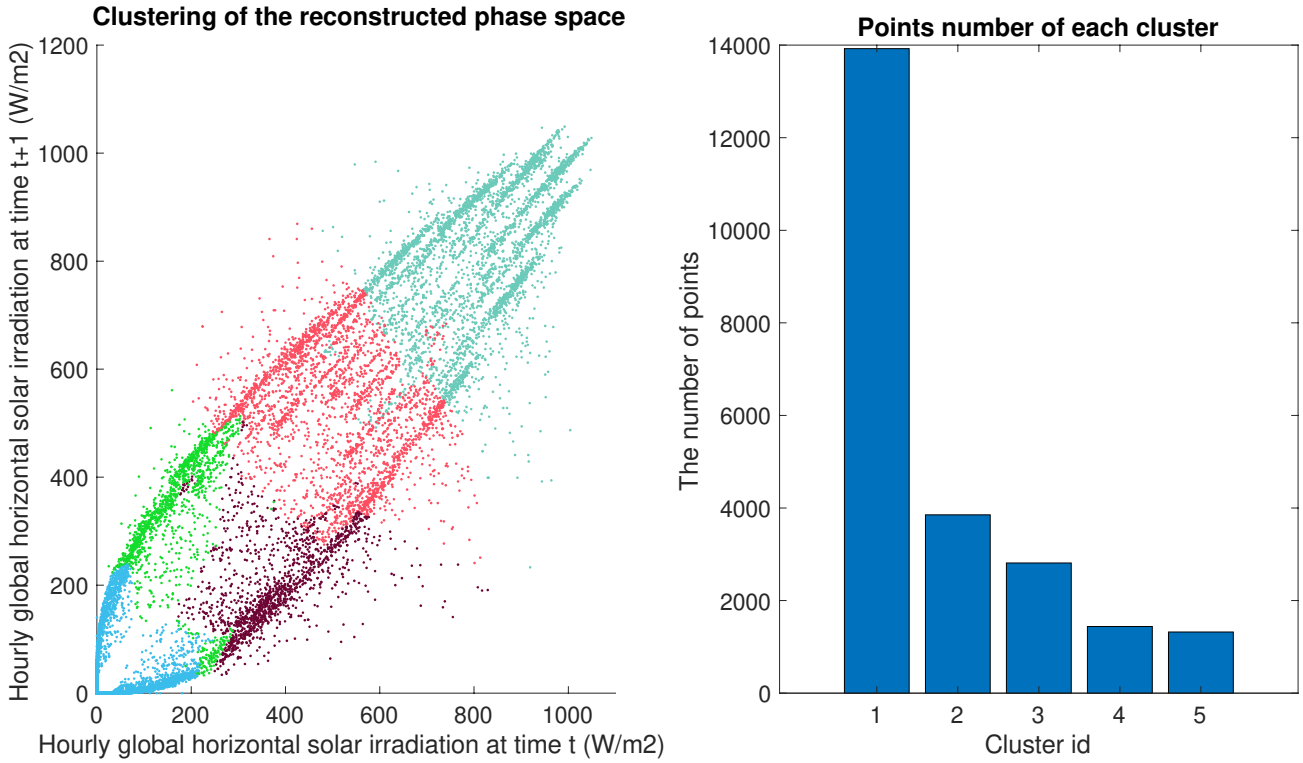

Fig.8 LGC for the reconstructed phase space of the training data (1st January 2018 to 30th September 2020, Ghardaia).

In addition, the comparison between measured and forecasted hourly solar irradiation data for three months as well as the measured data versus forecasted data is shown in Fig.9.From these figures, we can clearly see the goodness of the LGC- GMDH model. It forecasts in a good manner even on cloudy days with an R^2 value is equals to 98.57%. These results clearly prove the robustness of the proposed model to forecast solar radiation time series.

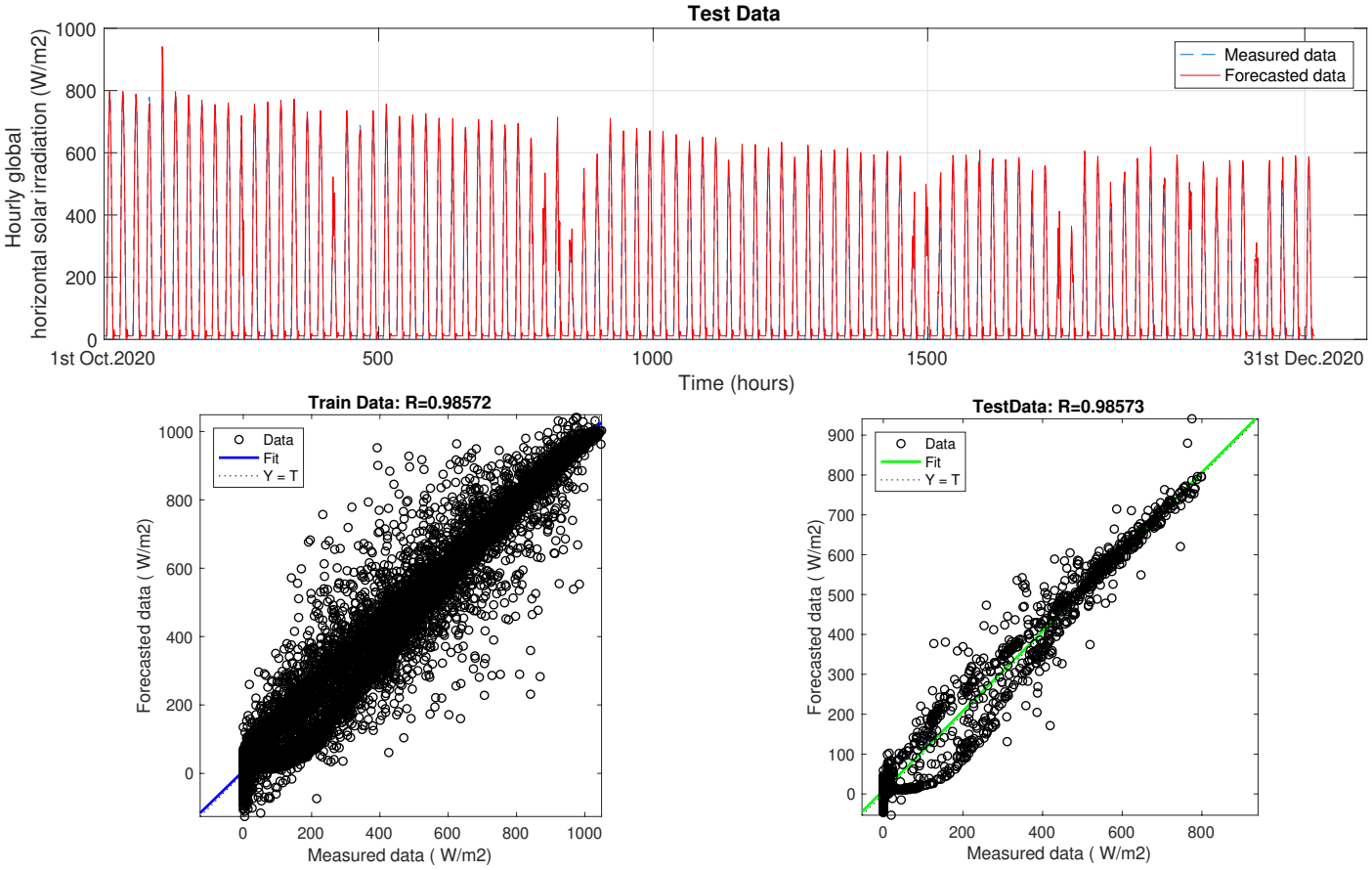

Fig.9 The measured versus forecasted test data LGC with GMDH model (1st Obtober 2020 to 31th December 2020, Ghardaia).

6. Comparison

Various comparison tests were conducted with other models to check the accuracy of the proposed LGC-GMDH forecast model. Single forecasting methods such as Artificial Neural Networks (ANN) (Faceira et al. 2015; Voyant et al. 2017), Support Vector Machines (SVM) (Wang et al. 2019), Recurrent Neural Networks (RNN) (Zhang and Behera 2012), and their hybrid models have been extensively utilized in solar irradiation forecasting. As a result, the use of single benchmark models, such as those presented by back propagation BP (Laopaiboon et al. 2019), SVM, generic RNN, and LSTM (Malakar et al. 2021; Kumari and Toshniwal 2021), is studied extensively in this paper. New multivariate time series models have also been tested alongside the three models mentioned above, which use multivariate time series data: the BP-MLP model, the RNN-MLP model, the LSTM-MLP model , and WPD-CNN-LSTM-MLP (Huang et al. 2021) . Moreover, clustered ANFIS network using fuzzy c-means, subtractive clustering,

615 and grid partitioning model presented in (Benmouiza and Cheknane 2018) , which provides good results compared to
616 other models is also considered in this comparison.

617 For the comparison purpose, solar irradiation data and other related climatic parameters (including air
618 temperature, relative humidity, wind speed, and other data) for the location of Denver (39440 N, 105 and 110 W,
619 Colorado, USA) in the United States are utilized to assess the performance of the proposed model. The data was taken
620 from the Measurement and Instrumentation Data Center (MIDC), National Solar Irradiation Database supplied by the
621 National Renewable Energy Laboratory, https://midcdmz.nrel.gov/. The dataset from 1st January 2012 to 31st
622 December 2015, was utilized as the training dataset, and the dataset from 1st January 2016 to 31st December 2016 was
623 used for testing the models. The original data is captured and recorded on a continuous minute basis ,then this data
624 was transformed to hourly data.

625 The RMSE , MAE , R^2 and FS have been chosen as error metrics in order to judge the goodness of our proposed
626 model. The results are expressed in Table 2 .

627 **Table 2** Comparison of RMSE , MAE , R^2 and FS for Denver (January 1, 2016 to December 31, 2016).

	Model	RMSE	MAE	R^2	FS
Single models	BP	76.9272	31.5464	0.9597	0.3167
	SVM	77.9152	52.8904	0.9599	0.3086
	RNN	70.5922	35.6135	0.9666	0.3730
	LTSM	68.7213	30.6064	0.9680	0.3896
Hybrid models	BP-MLP	63.1678	25.4427	0.9730	0.4389
	RNN-MLP	59.8450	23.1430	0.9766	0.4684
	LSTM-MLP	50.5665	19.5735	0.9832	0.5513
	WPD-CNN-LSTM-MLP	46.1336	20.3853	0.9858	0.590
Clustering models	ANFIS-clustering	45.5463	19.2985	0.9860	0.598
	LGC-GMDH	43.7268	19.1856	0.9887	0.613

628

629 Moreover, the errors bars for these results are shown in Fig.10.

630

631

632

633

634

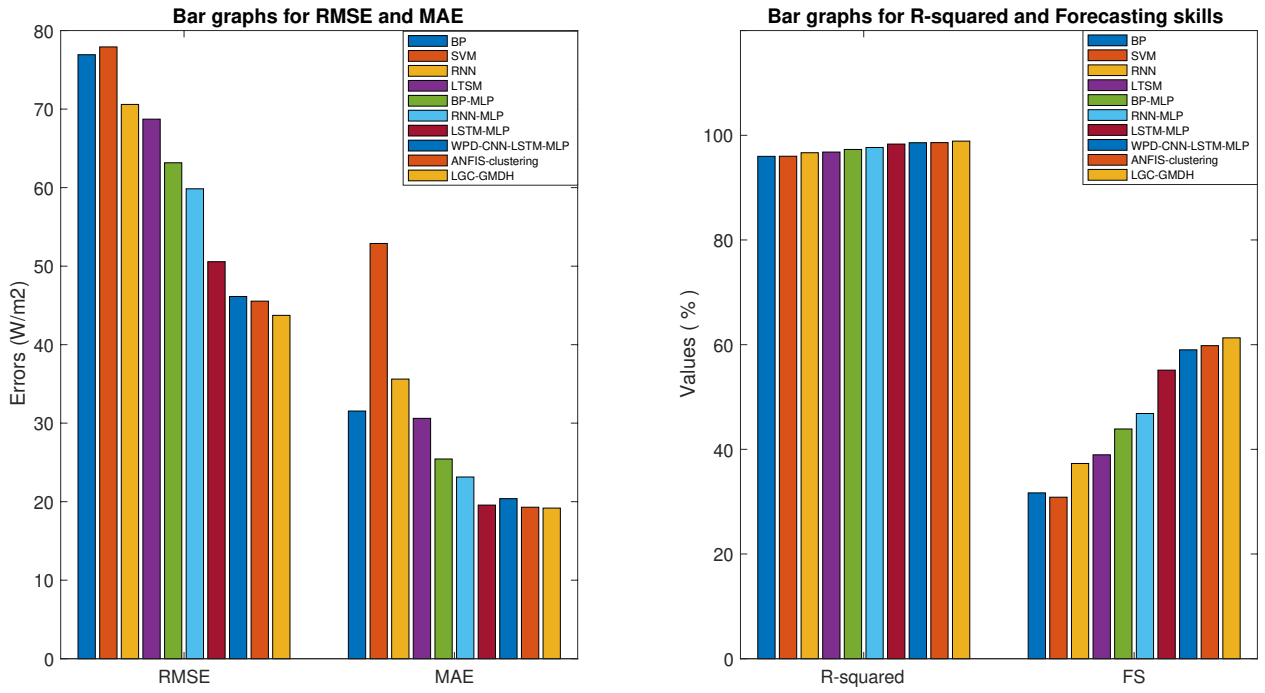

Fig. 10 Performance evaluation of different models for the region of Denver ,1st January 2016 to 31st December 2016 using BP, SVM , RNN , LTSM , BP-MLP, RNN-MLP- LSTM-MLP, WPD-CNN-LTSM-MLP, ANFIS clustering and LGC-GMDH.

As can be seen, the prediction accuracy of single models, such as BP, SVM, RNN, and LSTM, is not favourable. However, the prediction error is significantly reduced once the models are integrated with the other network structure. The lowest RMSE values for single models equal 68.7213 W/m^2 for LTSM, and a FS equals to 0.3167 for the BP model. In comparison, these errors are much lower in hybrid models with an RMSE equal to 46.1336 W/m^2 for WPD-CNN-LSTM-MLP with a FS equal to 0.590. Furthermore, the category of clustering hybrid models gives more accuracy. The errors have been improved in the ANFIS-clustering model, which, RMSE, MAE, R^2, and FS are 45.5463 W/m^2, 19.2985 W/m^2, 98.60% and 59.8%, respectively. However, our proposed LGC-GMDH model has better performance than all mentioned ones, with an RMSE, MAE, R^2, and FS of 43.7268 W/m^2, 19.1856 W/m^2, 98.87% and 61.3%, respectively. This comparison concludes that the LGC-GMDH can be used to forecast the hourly global solar irradiation data, with condition to choose the best configuration for better forecasting performance.

7. Conclusion

This paper suggested a new hybrid LGC-GMDH machine learning model for 1-hour horizon global horizontal irradiation forecasting. In this model, we have developed a sophisticated three-phase structure. The data were divided into training and testing data sets using k-cross validation. The clustering using the local gravity algorithm was used in order to get groups with similar dynamic characteristics. To this end, a phase space reconstruction was used in order to build a 2-dimension representation of the global horizontal irradiation data. The obtained clusters were used as inputs to the GMDH network in order to obtain the local forecasts. Then, a global forecast is then reconstructing. The proposed model achieved an excellent forecasting result because it takes into consideration the frequency characteristics of the irradiation time series and the power of the neural networks.

Comparing it with other models, the proposed model shows the lowest RMSE error equals 43.7268 W/m2, and an R^2 equals 0.9887, and the highest FS equals 0.613. It also illustrates that network structure has a significant impact on the model's overall accuracy. The LGC-GMDH deep learning model has shown that they provide accurate hour-ahead irradiation forecasts. There will be enough information to find the correct references for practical applications. Future work can test optimisation methods for optimal network configuration choosing to refine results further.

8. Declarations

Data availability Data used within this research are available upon request from the corresponding author.

Ethics approval No ethical approval is foreseen since the research described by this paper only concerns the analysis of meteorological data that have been made public by the laboratory that collected them and that no participant has been involved not directly nor indirectly as the subject of the research itself.

Consent to participate No participant has been involved not directly nor indirectly in the research itself.

Consent for publication All the authors mentioned in the manuscript have agreed to authorship, read, and approved the manuscript, and given consent for submission and subsequent publication of the manuscript. All named authors have agreed with the order of authorship before submission, and all authors have agreed to the name of the corresponding author. The manuscript has not been published anywhere else.

Competing interests The authors declare no competing interests.

Funding No funds, grants, or other support was received to assist with the preparation of this manuscript.

Author contributions The corresponding author [BENMOUIZA Khalil] contributed to the study conception and design. In addition, he performed the material preparation, data collection and analysis. The first draft of the manuscript was written by [BENMOUIZA Khalil]. Also, he read and approved the final manuscript.

9. References

Ahmed R, Sreeram V, Mishra Y, Arif MD (2020) A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew Sustain Energy Rev 124:109792. https://doi.org/10.1016/J.RSER.2020.109792

Alkhayat G, Mehmood R (2021) A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI 4:100060. https://doi.org/10.1016/J.EGYAI.2021.100060

Álvarez-Alvarado JM, Ríos-Moreno JG, Obregón-Biosca SA, et al (2021) Hybrid Techniques to Predict Solar Radiation Using Support Vector Machine and Search Optimization Algorithms: A Review. Appl Sci 2021, Vol 11, Page 1044 11:1044. https://doi.org/10.3390/APP11031044

Auvergne M, M. (1988) Singular value analysis applied to phase space reconstruction of pulsating stars. A&A 204:341–348

Benali L, Notton G, Fouilloy A, et al (2019) Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. Renew Energy 132:871–884. https://doi.org/10.1016/J.RENENE.2018.08.044

Benmouiza Khalil (2015) Quantification of solar radiation in Algeria, application to the sizing of photovoltaic systems. University of Tlemcen

Benmouiza K, Cheknane A (2013) Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models. Energy Convers Manag 75:. https://doi.org/10.1016/j.enconman.2013.07.003

Benmouiza K, Cheknane A (2018) Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting. Theor Appl Climatol 1–13. https://doi.org/10.1007/s00704-018-2576-4

Blanc P, Remund J, Vallance L (2017) Short-term solar power forecasting based on satellite images. Renew Energy Forecast From Model to Appl 179–198. https://doi.org/10.1016/B978-0-08-100504-0.00006-8

Botchkarev A (2019) A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdiscip J Information, Knowledge, Manag 14:45–76. https://doi.org/10.28945/4184

Burianek T, Misak S (2016) Solar irradiance forecasting model based on extreme learning machine. EEEIC 2016 - Int Conf Environ Electr Eng. https://doi.org/10.1109/EEEIC.2016.7555445

C. O (2008) Design of hybrid differential evolution and group method of data handling networks for modeling and prediction. Inf Sci Comput Sci Intell Syst Appl An Int J 178:3616–3634. https://doi.org/10.1016/J.INS.2008.05.013

Caldas M, Alonso-Suárez R (2019) Very short-term solar irradiance forecast using all-sky imaging and real-time irradiance measurements. Renew Energy 143:1643–1658. https://doi.org/10.1016/J.RENENE.2019.05.069

Chu Y, Li M, Coimbra CFM (2016) Sun-tracking imaging system for intra-hour DNI forecasts. Renew Energy 96:792–799. https://doi.org/10.1016/J.RENENE.2016.05.041

DT K, L G (1992) Direct test for determinism in a time series. Phys Rev Lett 68:427–430. https://doi.org/10.1103/PHYSREVLETT.68.427

Faceira J, Afonso P, Salgado P (2015) Prediction of Solar Radiation Using Artificial Neural Networks. Lect Notes Electr Eng 321 LNEE:397–406. https://doi.org/10.1007/978-3-319-10380-8_38

Farlow SJ (1981) The GMDH Algorithm of Ivakhnenko. Am Stat 35:210. https://doi.org/10.2307/2683292

Fraser A, Swinney H (1986) Independent coordinates for strange attractors from mutual information. Phys Rev A, Gen Phys 33:1134–1140

Gan M, Huang Y, Ding M, et al (2012) Testing for nonlinearity in solar radiation time series by a fast surrogate data test method. Sol Energy 86:2893–2896. https://doi.org/10.1016/j.solener.2012.04.021

Ghayekhloo M, Ghofrani M, Menhaj MB, Azimi R (2015) A novel clustering approach for short-term solar radiation forecasting. Sol Energy 122:1371–1383. https://doi.org/10.1016/J.SOLENER.2015.10.053

Grassberger P, Procaccia I (1983) Characterization of Strange Attractors. Phys Rev Lett 50:346.

https://doi.org/10.1103/PhysRevLett.50.346

Guermoui M, Melgani F, Gairaa K, Mekhalfi ML (2020) A comprehensive review of hybrid models for solar radiation forecasting. J Clean Prod 258:120357. https://doi.org/10.1016/J.JCLEPRO.2020.120357

Huang X, Li Q, Tai Y, et al (2021) Hybrid deep neural model for hourly solar irradiance forecasting. Renew Energy 171:1041–1060. https://doi.org/10.1016/J.RENENE.2021.02.161

Kennel MB, Brown R, Abarbanel HDI (1992) Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys Rev A 45:3403–3411. https://doi.org/10.1103/PhysRevA.45.3403

Khalil B, Ali C (2016) Density-based spatial clustering of application with noise algorithm for the classification of solar radiation time series. In: 2016 8th International Conference on Modelling, Identification and Control (ICMIC). IEEE, pp 279–283

Klipp E, Herwig R, Kowald A, et al (2005) Systems Biology in Practice. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. 1137–1143

Kondo T (1998) GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem. Proc SICE Annu Conf 1143–1148. https://doi.org/10.1109/SICE.1998.742993

Kumari P, Toshniwal D (2021) Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. Appl Energy 295:117061. https://doi.org/10.1016/J.APENERGY.2021.117061

Lai CS, Zhong C, Pan K, et al (2021) A deep learning based hybrid method for hourly solar radiation forecasting. Expert Syst Appl 177:114941. https://doi.org/10.1016/J.ESWA.2021.114941

Laopaiboon T, Ongsakul W, Panyainkaew P, Sasidharan N (2019) Hour-Ahead Solar Forecasting Program Using Back Propagation Artificial Neural Network. Proc Conf Ind Commer Use Energy, ICUE 2018-October: https://doi.org/10.23919/ICUE-GESD.2018.8635756

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. The Regents

of the University of California

Malakar S, Goswami S, Ganguli B, et al (2021) Designing a long short-term network for short-term forecasting of global horizontal irradiance. SN Appl Sci 2021 34 3:1–15. https://doi.org/10.1007/S42452-021-04421-X

Mazorra Aguiar L, Pereira B, David M, et al (2015) Use of satellite data to improve solar radiation forecasting with Bayesian Artificial Neural Networks. Sol Energy 122:1309–1324. https://doi.org/10.1016/j.solener.2015.10.041

Michel O, Flandrin P (1996) Application of methods based on higher-order statistics for chaotic time series analysis. Signal Processing 53:133–148. https://doi.org/10.1016/0165-1684(96)00082-5

Nazerfard E, Shouraki SB, Hakami V (2006) Evolutionary GMDH-based Identification of Building Blocks for Binary-Coded Systems. 1900–1904. https://doi.org/10.1109/ICTTA.2006.1684679

Onwubolu G (2016) GMDH-Methodology and Implementation in MATLAB. GMDH-Methodology Implement MATLAB. https://doi.org/10.1142/P982

Packard NH, Crutchfield JP, Farmer JD, Shaw RS (1980) Geometry from a Time Series. Phys Rev Lett 45:712. https://doi.org/10.1103/PhysRevLett.45.712

Perez R, Lorenz E, Pelland S, et al (2013) Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Sol Energy 94:305–326. https://doi.org/10.1016/J.SOLENER.2013.05.005

Premalatha N, Valan Arasu A (2016) Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms. J Appl Res Technol 14:206–214. https://doi.org/10.1016/J.JART.2016.05.001

Refaeilzadeh P, Tang L, Liu H (2009) Cross-Validation. Encycl Database Syst 532–538. https://doi.org/10.1007/978-0-387-39940-9_565

Schmidt T, Kalisch J, Lorenz E, Heinemann D (2016) Evaluating the spatio-temporal performance of sky-imager-based solar irradiance analysis and forecasts. Atmos Chem Phys 16:3399–3412. https://doi.org/10.5194/acp-16-3399-2016

Schüler D, Wilbert S, Geuder N, et al (2016) The enerMENA meteorological network – Solar radiation

measurements in the MENA region. AIP Conf Proc 1734:150008. https://doi.org/10.1063/1.4949240

Shadab A, Said S, Ahmad S (2019) Box–Jenkins multiplicative ARIMA modeling for prediction of solar radiation: a case study. Int J Energy Water Resour 2019 34 3:305–318. https://doi.org/10.1007/S42108-019-00037-5

Soubdhan T, Ndong J, Ould-Baba H, Do MT (2016) A robust forecasting framework based on the Kalman filtering approach with a twofold parameter tuning procedure: Application to solar and photovoltaic prediction. Sol Energy 131:246–259. https://doi.org/10.1016/J.SOLENER.2016.02.036

Tadj M, Benmouiza K, Cheknane A, Silvestre S (2014) Improving the performance of PV systems by faults detection using GISTEL approach. Energy Convers Manag 80:298–304. https://doi.org/10.1016/J.ENCONMAN.2014.01.030

Takens F (1981) Detecting strange attractors in turbulence. 366–381. https://doi.org/10.1007/BFB0091924

Vaishnav V, Vajpai J (2018) Seasonal Time Series Forecasting by Group Method of Data Handling. 2018 IEEE Int Students' Conf Electr Electron Comput Sci SCEECS 2018. https://doi.org/10.1109/SCEECS.2018.8546886

Verbois H, Huva R, Rusydi A, Walsh W (2018) Solar irradiance forecasting in the tropics using numerical weather prediction and statistical learning. Sol Energy 162:265–277. https://doi.org/10.1016/J.SOLENER.2018.01.007

Vindel JM, Polo J (2014) Markov processes and Zipf's law in daily solar irradiation at earth's surface. J Atmos Solar-Terrestrial Phys 107:42–47. https://doi.org/10.1016/J.JASTP.2013.10.017

Voyant C, Muselli M, Paoli C, Nivet ML (2012) Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. Energy 39:341–355. https://doi.org/10.1016/j.energy.2012.01.006

Voyant C, Notton G, Kalogirou S, et al (2017) Machine learning methods for solar radiation forecasting: A review. Renew Energy 105:569–582. https://doi.org/10.1016/J.RENENE.2016.12.095

W.M.O. (1981) Meteorological aspects of the utilization of solar radiation as an energy source, illustrate. Secretariat of the World Meteorological Organization

Wang B, Che J, Wang B, Feng S (2019) A Solar Power Prediction Using Support Vector Machines Based on Multi-source Data Fusion. 2018 Int Conf Power Syst Technol POWERCON 2018 - Proc 4573–4577. https://doi.org/10.1109/POWERCON.2018.8601672

Wang Z, Yu Z, Philip Chen CL, et al (2018) Clustering by Local Gravitation. IEEE Trans Cybern 48:1383–1396. https://doi.org/10.1109/TCYB.2017.2695218

Water PR, Kerckhoffs EJH, Van Welden D (2000) GMDH-based dependency modeling in the identification of dynamic systems. In: Proceedings of the 14th European Simulation Multiconference on Simulation and Modelling: Enablers for a Better Quality of Life. Society for Computer Simulation International, San Diego, pp 211–218

Whitney H (1936) Differentiable manifold. Ann Math 37:645–680. https://doi.org/10.2307/1968482

Wong TT (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit 48:2839–2846. https://doi.org/10.1016/J.PATCOG.2015.03.009

Yadav AP, Behera L (2014) Solar Radiation forecasting using neural networks and Wavelet Transform. IFAC Proc Vol 47:890–896. https://doi.org/10.3182/20140313-3-IN-3024.00218

Zhang N, Behera PK (2012) Solar radiation prediction based on recurrent neural networks trained by Levenberg-Marquardt backpropagation learning algorithm. 2012 IEEE PES Innov Smart Grid Technol ISGT 2012. https://doi.org/10.1109/ISGT.2012.6175757