Journal of Cleaner Production 279 (2021) 123708

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

A novel approach to predict CO_2 emission in the agriculture sector of Iran based on Inclusive Multiple Model

Elham Shabani [a,*], Babollah Hayati [a], Esmaeil Pishbahar [a], Mohammad Ali Ghorbani [b,c], Mohammad Ghahremanzadeh [a]

[a] Department of Agriculture Economics, University of Tabriz, Iran
[b] Water Engineering Department, University of Tabriz, Iran
[c] Department of Civil Engineering, Istanbul Technical University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 14 January 2020
Received in revised form
4 July 2020
Accepted 8 August 2020
Available online 17 August 2020

Handling editor is Cecilia Maria Villas Bôas de Almeida

Keywords:
Artificial neural network
CO_2 emission
Gaussian process regression
Inclusive multiple model
Prediction

ABSTRACT

Due to the significant effects of CO_2 emissions on climate change and global warming, as well as its serious hazards to human health, the prediction of CO_2 emission is a vital issue. The main aim of this paper is to evaluate the power of the Inclusive Multiple Model (IMM) as a novel approach to predict CO_2 emission in the agriculture sector of Iran. For the same, we implemented the environmental Kuznets curve (EKC) specification and data from 2003 to 2017 for 28 provinces of Iran. In the first level, various specifications were implemented for each of the Multiple Regression (MLR), Gaussian Process Regression (GPR), and Artificial Neural Network (ANN) models. In the second level, an IMM model was implemented for treating the outputs of the best specification out of the MLR, GPR, and ANN models as inputs to an ANN model. The performance of the models was compared with the Taylor diagram and innovation and unique graphs. Findings indicated that the IMM model with $CC = 0.81$, $RMSE = 0.69$, the highest residuals between -5 and 5 (37.84%), and the lowest distance from observation points (1.857) estimated CO_2 emission values more precisely. These improvements indicate that there are possible directions for future research activities. Due to the most accuracy of the IMM, it is recommended to use this method to predict CO_2 emission to adopt appropriate policies for reducing air pollution.

1. Introduction

One of the most essential principles for sustainable growth & development and a healthy life for a human is a healthy environment. Environmental pollution affects the quality as well as the natural cycle and causes harmful effects on humans, animals, plants, etc. The widespread use of fossil fuels in various economic activities, such as agriculture, leads to the release of toxic and dangerous gases such as carbon dioxide (CO_2) that plays an important role in global warming and climate change as well as in human activities including industrial productivity, agriculture produces, immigration, and population (Fang et al., 2018). Accordingly, the issue of environmental pollution and climate change induced by greenhouse gas emission is a serious issue in the present world and climate change scenario (Zhang et al., 2019).

CO_2 is the major greenhouse gas that has increased by about 30% since the beginning of the pre-industrial era around 1750. Global carbon emissions in 2018 were 37.1billion tones. Iran, with 672 million tons of CO_2 emission in 2018, has been ranked 7th among the most carbon-intensive countries worldwide (Global carbon atlas, 2018). In Iran, 48.2% of GDP is equivalent to 34 thousand billion USD is spent on pollution caused by air.

In addition to being the engine of growth and development, the agricultural sector activities are also a source of air pollution. According to the World Resources Institute, the agricultural sector earmarked 6 billion tons of greenhouse gases, equivalent to 13 percent of the total greenhouse gas emissions in 2014. Iran's economy consists of four main sectors, including agriculture, manufacturing and mining, oil, and services. It is necessary to address the issue of CO_2 emission in the agricultural sector, because among the polluting gases, CO_2 with 12.5 million tones plays a major role in air pollution by the agriculture sector. Besides, in 2014, in the agricultural sector of Iran, CO_2 has the highest emission

* Corresponding author.
E-mail addresses: Elhamshabani@tabrizu.ac.ir (E. Shabani), b-hayati@tabrizu.ac.ir (B. Hayati), pishbahar@yahoo.com (E. Pishbahar), ghorbani@tabrizu.ac.ir (M.A. Ghorbani), ghahremanzadeh@tabrizu.ac.ir (M. Ghahremanzadeh).

https://doi.org/10.1016/j.jclepro.2020.123708
0959-6526/

intensity equal to 159.769 tons per billion. Also, it has the highest social costs equivalent to 100.9 billion rials, among other pollutants.

CO_2 emission in the world especially in Iran is a serious problem, and it is very vital to pay attention to this issue and to provide appropriate solutions to reduce the emission of this pollutant because the release of CO_2 is a serious threat to the health of humans and other living organisms, climate change, and global warming. The first step in planning and programming to reduce CO_2 emission is to be aware of the concentration of CO_2 in the future. In this regard, it is essential to choose the best model that predicts a variable with the lowest errors because the accuracy of a method is particularly important in the prediction process. A prediction is a powerful tool for programming. Nowadays, the prediction of CO_2 emission is a fundamental key for energy programming, particularly those related to low carbon. On the other side, it is important to study the prediction of CO_2 emissions for proper planning to achieve economic objectives and programs for sustainable growth.

Due to the serious dangers of CO_2 emission to human health and natural ecosystems, various studies have predicted the CO_2 emission as an important policy tool. The main issue in these studies has been the development of different new models and evaluate their predictive accuracy based on evaluation criteria. So, these studies, while using traditional methods such as Multiple Regression (MLR) and Autoregressive integrated moving average (ARIMA), have sought to develop new models with higher predictive accuracy. In this regard, several smart models and algorithms have been developed, each of which has its advantages and disadvantages. Here are some recent studies on predicting carbon emissions using a variety of methods.

Behrang et al. (2011) used the Bees algorithm and Artificial Neural Network (ANN) to forecast world CO_2 emissions. Lin et al. (2011) used the grey forecasting model to estimate future CO_2 emissions in Taiwan. Pao et al. (2012) forecasted CO_2 emission in China using ARIMA, grey model, and nonlinear grey Bernoulli model (NGBM). They concluded that NGBM model with low RMSE, MAE, and MAPE was the best model for forecasting CO_2 emission. Abdelfatah et al. (2013) implemented the Adoptive Particle Swarm Optimization (APSO) algorithm for forecasting global CO_2 emissions, and results confirmed the high accuracy of the developed model over the other examined models. Lotfalipour et al. (2013) predicted CO_2 emissions in Iran based on the Grey System and the ARIMA model, and results of the RMSE, MAE, and MAPE showed that the Grey system was more accurate in forecasting than the other methods of prediction.

Samsami (2013) used data from 1981 to 2009, and Ant Colony Optimization (ACO) algorithm to forecast CO_2 emission in Iran. The high accuracy of the ACO algorithm in this study and similar studies with a small number of observations justifies the use of this algorithm in studies with a limited number of observations. Taghavifar et al. (2015) used the ANN model for the prediction of pollutants from n-heptane fueled engines. Noori et al. (2015) predicted CO_2 emission with Adoptive Neuro Fuzzy Inference Systems (ANFIS), ANN, and Support Vector Regression (SVR) models and selected the SVR as the best model according to model selection criteria. Saleh et al. (2015) used the Support Vector Machines (SVM) model for the prediction of CO_2 emission. Sun and Liu (2016) used the least squares support vector machine (LSSVM) for the prediction of three major industries and residential consumption CO_2 emissions in China.

Ann et al. (2017) used the recurrent neural network of deep learning for the prediction of CO_2 emission and gained the highest accuracy; however, using this method requires a large number of observations. Yu et al. (2017) implemented the Grey system model to predict CO_2 emission in China. Libo et al. (2017) used Multiple Linear regression for the prediction of CO_2 emission. Chen et al. (2018) used four machine learning algorithms, including SVM, backpropagation neural network (BPNN), Gaussian processes (GP), and M5P to predict CO_2 and concluded that the SVM was the best model according to performance metrics. Kardani et al. (2018) showed the modeling of CO_2 absorption in polyionic liquids utilized radial basis function artificial neural network (RBFANN) and LSSVM combined with group contribution (GC) method, RBFANN-GC and LSSVM-GC, respectively, and compare with multi-layer perceptron artificial neural network (MLPANN) and ANFIS. Error analyses indicate the great performance of both RBFANN-GC, and LSSVM-GC in the estimation of CO_2 absorption in polyionic liquids.

For the prediction of CO_2 emission in India, Sangeetha and Amudha (2018) used the MLR and PSO algorithm based on the nonlinear model, and reported that the PSO model could achieve a highly accurate estimation compared to the MLR model. Fang et al. (2018) used an improved Gaussian Process Regression based on the PSO algorithm for forecasting CO_2 emission in the USA, China, and Japan. Recently, Xu et al. (2019) have used dynamic nonlinear ANN approach and scenario analysis to predict CO_2 emissions in China. The results of this study and similar studies confirm the high accuracy of nonlinear models compared to linear models in predicting carbon emissions. Wang and Li (2019) have recently forecasted CO_2 emission in China with the PSO algorithm and the grey Verhulst model using data from 1990 to 2014.They concluded that the PSO-DNE grey Verhulst model to be the best model according to model selection criteria. These results demonstrate the effectiveness of the PSO algorithm in studies with a limited number of observations. Due to the review of studies and the high accuracy of new smart models, there is a possibility of new routes and new models to increase the accuracy of prediction.

Prediction is a part of our desire to control the future and to imagine the future that we want. Before the implementation of any pollution reduction policy, to adopt the appropriate policy and evaluate the effectiveness of the applied policy, the policymaker needs to be aware of the status of pollutants in the case of any pollution reduction policy. In this regard, a prediction is an important tool for environmental policymakers. The main goal in the prediction is to choose a method with high accuracy. Inclusive Multiple Model (IMM) was first introduced implicitly by Khatibi et al. (2017) for streamflow prediction in the Bear River of the USA, and then explicitly by Khatibi et al. (2020). The underlying methodology is implicit in quite a few of the papers recently published by him through various groups of modeling research.

The latter publication provides greater detail for the rationale of IMM in terms of a set of activities involving: 1-Model Reuse (MR), 2-Hierarchy and,or Recursion (HR), 3-A provision of 'Elastic' model-Learning Environment (ELE) 4-Goal-Orientation (GO)—leading to the acronym of RHEO. The above are four dimensions forming a framework for the implementation of IMM. This is just an explanatory framework to integrate a wide range of modeling activities. The underlying issue is that IMM provides a shift from the traditional section of the a superior model to learning from models to ensure more enhanced accuracy. This paper is one such application. In practice, the results of the studies based on the IMM method in other scientific fields have shown better performance and more accuracy of the IMM model than the other examined models. Because the use of the IMM ensures the collection of adequate information hence there is no potential capacity of residuals for extracting further information.

So, as a model with high accuracy would be a strong tool for policymakers, planners, and institutes with concerns about hazardous consequences of air pollution, and to be aware of carbon emission trends to adopt appropriate policies, including incentives and taxation to control air pollution, developing a more accurate model is an essential issue. Despite the high accuracy and extensive advantages of the IMM, there is no study that has used the IMM to predict CO_2 emission, and this is the first study in the field of CO_2 prediction by the new IMM method, which is the innovation of this paper. Besides, the present paper used various novel graphs to evaluate the power of the models. We developed an innovative evaluation criterion based on the Taylor diagram, which takes into account three important criteria in evaluating the accuracy of a model. For each implemented model, shows a measure that indicates the distance between the values predicted by the model and the actual values, which it reduces the error of decision making.

The paper progresses as follows. Section 2 involves materials and method, including study area and data description, implemented models, and performance metrics. Section 3 includes results and discussion, and Section 4 presents conclusions.

2. Materials and methods

2.1. Study area and data description

The literature of the environmental Kuznets curve (EKC), which is used to determine the relationship between environmental quality and economic growth (Wang and Li, 2019), was used in this study to determine variables affecting CO_2 emission. According to EKC, the environmental damage first increases by increasing per capita income as a factor of economic growth. Still, it will decrease after a “turning point”, showing the per capita income of the developing countries. There is, therefore, an interval U-shaped relationship between environmental quality and economic growth. EKC first began with a paper by Grossman and Kruger in 1992, and since then, most studies have focused on EKC. A bulk of such studies have used CO_2 emissions as a measure of environmental quality (for example, Jalil and Mahmud, 2009; Tien Pao and Tsai, 2011; Alam et al., 2016; Ben Jebli et al., 2016; Wang et al., 2017).

In general, there are three different types of literature on environmental pollution and economic growth. The first group of studies explores the relationship between economic growth and environmental pollution in the form of standard EKC only with per capita GDP and GDP square (Chen et al., 2018; Galeotti et al., 2009; Lee and Lee, 2009). The second group believes that fossil fuels are the main cause of CO_2 emissions (Erol and Yu, 1987; Apergis and Payne, 2009). Some studies have looked at both energy consumption and economic growth (Ozcan, 2013; Kasman and Duman, 2015; Aydin and Esen, 2018; Pata, 2018; He and Lin, 2019). According to this literature, three critic variables influencing CO_2 emissions are per capita GDP, per capita GDP square, and energy consumption. Besides, another factor is income distribution which is often neglected in EKC evaluation, but it should be considered as an effectively predicted variable to avoid omitted variable bias (Boyce, 1994; Magnani, 2000; Borghesi, 2000; Golley and Meng, 2012; Hao et al., 2016; Ridzuan, 2019; Uzar and Eyuboglu, 2019).

In the present study to predict the CO_2 emission in the Iranian agriculture sector, four different models are implemented by considering GDP, GDP^2, energy consumption, and GINI index variables. GDP is a measure of economic growth and is the Gross Domestic Products in the agriculture sector of Iran that divided into the agriculture population to earn per capita GDP in the agriculture sector. GDP^2 is the square of the per capita GDP, Energy consumption is the amount of the fossil fuels that are used with agriculture sector of Iran, and GINI index is the measure of income inequality in the agriculture sector of Iran that ranges are between 0 (unequal) to 100 (equal).

Iran is located in southwestern Asia and the Middle East region, with an area of about 1.6 million square kilometers. It is one of the largest countries in the region with a population of about 79 million people. Iran is the 7th biggest CO_2 emitter in the world. The agriculture sector is one of the critical sectors in the country, with CO_2 being the highest pollutant in this sector. The required data, including CO_2 emission and energy consumption in the agriculture sector of Iran, were collected from the energy balance sheets available on the website of the Ministry of energy of Iran. Data related to GDP, population, producer price indices, and GINI index in the agriculture sector of Iran were extracted from the regional accounts of the website of the Statistics Center of Iran. Real GDP is obtained by dividing nominal values by the producer price index, and per capita values were earned by dividing total values to the population. Required data were collected for 28 provinces during 2003–2017. Table 1 present the statistical characteristics of the data.

2.2. Implemented models

This study consists of two main levels. The first level predicts CO_2 emission with MLR, GPR, and ANN models with various structures and selects the best models using performance metrics. At the second level, an MMS model is used for treating the outputs of the MLR, GPR, and ANN as inputs to an ANN model. Then, the whole estimated models are compared with performance metrics to evaluate the potential performance of the new IMM model. We used Wolfram Mathematica 11.3. software to run the models and plotted the diagrams.

2.2.1. Multiple regression (MLR)

First implemented by Pearson (1908), multiple regression (MLR) is a statistical and powerful method that implements some explanatory variables to predict the outcome of a dependent variable (Pearson and Lee, 1908). The main goal of multiple regression is to comprehend more relationships among some independent variables and a dependent variable. Multiple regression can be expressed as:

$$Y = f(X\,;\beta) + \varepsilon \tag{1}$$

where Y is the dependent or predicted variable, X is a vector of independent or predictor variables, β is a vector of the coefficient of predictor variables, ε is the error term, and f is the functional form that could be linear or non-linear (logarithmic, quadratic, cubic, exponential etc.). It is so important and a considerable issue to determine the correct functional form to present the relationship between variables. While a linear regression will turn into a curve using a polynomial term—a quadratic or cubic term, it yet certifies a linear model because it is X that is squared or cubed, not the coefficient β. It can show a straightforward way to model curves without having to model embroiled non-linear models. In this study, two linear and nonlinear (quadratic) functional forms are used to predict CO_2 emissions (see Libo et al., 2017). To predict CO_2 emissions with the MLR model, equation (2) was used:

$$CO_2 = \beta_1 GDP + \beta_2 GDP^2 + \beta_3 EC + \beta_4 GINI + \varepsilon \tag{2}$$

2.2.2. Artificial neural network (ANN)

ANN is a supervised machine learning method. The first attempts at simulation using a logic model were made by Warren Mc Calk and Walter Pitts in the early 1940s, which today constitute the main building block of most ANNs. The basic idea behind these networks is based on biological neural networks. This model can be useful for problems that are easily handled by the human mind. In fact, the way the mind works is kind of an inspiration for a model

Table 1
Statistical characteristics of study variables.

Variables	Variable define	Unit	N. of data	Mean	Standard deviation	Skewness	Max.	Min.
CO_2	Per capita CO_2	Metrics kg	420	5.766	1.618	0.731	12.132	2.775
GDP	Real Per capita GDP	Billion Rial	420	0.519	0.602	0.110	10.434	0.046
GDP^2	Square of GDP	Billion Rial	420	631.441	533.6	0.199	108,321	2.163
EC	Energy use	1000 barrels of oil equivalent	420	1017.899	446.73	1.144	3351.71	301.92
GINI	Rural GINI index	–	420	34.150	3.870	−0.075	44.74	24.610

for brain-like capabilities. ANN is the most popular and acceptable model for predicting a wide range of variables, such as pollutants. The advantages of an ANN model are generalizability, nonlinear calculations, communicate one input and one output, information retrieval, adaptability, response to noise data, error tolerance, and learning.

There are two types of ANN models, including Multilayer Perceptorn (MLP) and Radial Basis Function (RBF). MLP is one of the basic artificial models that simulate the transformation function of the human brain. There are three types of neuron layers in an MLP model: i) the input layer that detects the input and sends the input signal to the next layer in terms of its powerful connection with the next layer; ii) the hidden layers to weight the inputs and the relationship between them; and iii) the output layer shows the decision output (Oliveira et al., 2018). The general form of an MLP-ANN could be shown as:

$$Z = \nu_0 + \sum_{j=1}^{N_{hd}} \nu_j f\left(\omega_{oj} + \sum_{i=1}^{N_{in}} \omega_{ji} X_i\right) \tag{3}$$

Where Z is a network output, that is, CO_2 emission predicts, X_i is network input for input layer neuron i including *GDP, GDP^2, EC,* and *GINI*, N_{in} is the number of MLP inputs, the ω_{ji} is the weight of i input to j hidden layer neuron, *f()* is the activation function constant for all hidden neurons, N_{hd} is the number of hidden neurons, ν_j is the weight for output from hidden neuron j, and ω_{oj} , ν_0 are neuron biases. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. There is three common activation function including:

$$\begin{aligned} f(z) &= z \text{ Linear function} \\ f(z) &= \frac{e^z - e^{-z}}{e^z + e^{-z}} \text{ Hyperbolic tangent ranges from } -1 \textit{ to } 1 \\ f(z) &= \frac{1}{1 + e^{-z}} \text{ Sigmoid function ranges from } 0 \textit{ to } 1 \end{aligned} \tag{4}$$

To optimize and training the MLP structure we used the Back-Propagation algorithm. To minimize the difference between predicted value with the ANN model and actual values in the training phase, optimal values for weights of the neurons in the output layer (ω_i) and bias term (b_i) should be calculated. The error function equation could be represented as Zarei et al. (2019):

$$E = \sum_p E(\omega) = \sum_p \sum_i \left(r_i^p - O_i^{p,l}\right) \tag{5}$$

Where p presents the number of training data (315), $O_i^{p,l}$ shows the ith neuron from the lth output layer, and r_i^p refers ith the real output of pth training data.

The weights and biases that are transformed through the BP algorithm could be shown as:

$$\omega_{i,j}^{l-1,l}(k+1) = \omega_{i,j}^{l-1,l}(k) - \lambda \frac{\partial E}{\partial \omega_{i,j}^{l-1,l}} \text{ Weight} \tag{6}$$

$$b_j^l(k+1) = b_j^l(k) \lambda \frac{\partial E}{\partial b_j^l} \text{ Bias} \tag{7}$$

Where the rate of learning is denoted with λ, and k presents the number of iterations.

2.2.3. Gaussian Process Regression (GPR)

Gaussian process regression (GPR) model is a nonparametric kernel-based probabilistic and full Bayesian learning algorithm model that creates waves in the sphere of machine learning. The Bayesian approach deduces a probability distribution over all possible values, while most of the famous supervised machine learning algorithms acquire absolute values for each parameter in a function. GPR model can be implemented in varied applications such as multivariate regression, experiment design, and model approximation, and also to solve nonlinear regression problems (Girard et al., 2003; bib_citation_to_be_resolvedQuiñonero-Candela and Rasmussen, 2005; Rasmussen and Williams, 2006; Hu and Wang, 2015; Chen et al., 2018; Fang et al., 2018). Easy to implement, working well on small datasets, self-adaptive to enable superior parameter estimation, combining different machine learning duties involving model training, uncertainty estimation, and hyperparameter estimation are the advantages of GPR compared to other machine learning methods (Hu and Wang, 2015). When using GP, it is possible to add prior awareness and specifications about the shape of the model by choosing various kernel functions. In GPR, it is supposed that the output variable measurements of y can be presented as:

$$y = f(X(k)) + \varepsilon \tag{8}$$

where f shows the unknown nonlinear function to be modeled, ε shows noise with a Gaussian distribution and variance σ_n^2 and X show measurements of input variables. The space of functions has a prior probability explained as a Gaussian process owning mean m(x) and covariance function cov (x, x') (Rasmussen and Williams, 2006).

$$f(x) \sim GP(m(x), cov(x, x')) \tag{9}$$

For a specified exemplar of predictor variables x_*, the response variable is predicted implementing the predictive probability distribution p (y*|X, y, x*) with mean and variance:

$$\begin{aligned} \hat{y}_* &= m(x_*) + k_*^T \left(K + \sigma_n^2 I\right)^{-1} (y - m(x_*)) \\ \sigma_{y_*}^2 &= k_* + \sigma_n^2 - k_*^T \left(K + \sigma_n^2 I\right)^{-1} k_* \end{aligned} \tag{10}$$

where k_* shows a vector that has the ability to be explained by $[k_*]_i = \mathrm{cov}(x_i, x_*)$, $k_* = \mathrm{cov}(x_i, x_*)$, K shows a covariance matrix with elements $[K]_{i,j} = \text{cov}(x_i, x_j)$, and I represents the identity matrix. Thus, it can be seen that, unlike the traditional regression methods in which only the parameters are implemented to specify the model prediction, here, the model output belongs to the training dataset X, y.

For definite predictions, the accessible data are implemented to establish the parameters of the mean and covariance function. The predictive probability distribution is wholly determined to implement these parameters, usually referred to as hyperparameters. The values of hyperparameters can be achieved by maximizing the log-likelihood function of the training datasets (Rasmussen and Williams, 2006):

$$\log p(y|X) = -\frac{1}{2}y^T\left(K+\sigma_n^2 I\right)^{-1}y - \frac{1}{2}\log\left(\left|K+\sigma_n^2 I\right)\right| - \frac{n}{2}\log(2\pi) \tag{11}$$

where n is the number of training datasets (see Chen et al., 2018). In the present study, y denotes CO_2 emission, and X denotes *GDP, GDP*^{2}*, EC,* and *GINI* index. The number of training data or n is equal to 280.

2.2.4. Inclusive Multiple Model (IMM) driven by artificial neural network

In the present study, the ANN model is implemented in two phases: i) level 1 as a traditional tool to predict CO_2 emission, and ii) level 2 in the IMM model. Essential information about the ANN model in levels 1 and 2 is presented in Table 2. IMM is used to ensure the extraction of sufficient information and no potential capacity to residuals for extracting further information. For combining the best models of level 1, there are several strategies, each of which will show different results. The strategy of implementing multiple models used in this paper is MM-ANN (IMM drove by the ANN) (Ghorbani et al., 2018).

The individual modeling results of the MLR, GPR, and ANN were regarded as a lower order modeling results and treated as the input data to an upper-level ANN model to be adverted to an MMS model. The target values of this IMM model are defined in terms of observed values, now implemented in MLR, GPR, and ANN. As introduced above, the IMM model needs its structure to be defined, and this is also denominated by the trial, and error manner. IMM is used by dividing the input and target data into training and testing phases, and this method of using ANN models is understood as supervised training (Khatibi et al., 2017). The diagram of this process is shown in Fig. 1.

2.3. Performance metrics

The following three metrics were implemented to appraise the performance of the four models: (i) Correlation Coefficient (CC) or R that shows the correlation structure between the observed and forecasted values, and the higher its value, the lesser the deviation, (ii) Standard Deviation (SD), which measures the dispersion of the forecasted values relative to their mean and is calculated as the square root of variance. The lower the SD value, the better the model performance, (iii) Root Mean Squared Error (RMSE), which indicates the discrepancy between the observed and modeled values. The lower the RMSE value, the better the model performance, and (iv) Relative Error (RE), which is the ratio of the absolute error by the modeled value, indicating that how good a modeled value is relative to the size of the measured variable. It is a good reflection of the maximum error. The Taylor diagram (Taylor, 2001) is also used to evaluate the models, which is a single diagram and provides a visual representation of the CC, SD, and RMSE of the observed and predicted data. The Correlation Coefficient, RMSE, and SD are considered as follow Zarei et al. (2019):

$$R = CC = \left(1 - \sum_{i=1}^{N}\frac{\left(CO2_i^{actual} - CO2_i^{predicted}\right)}{\left(CO2_i^{actual} - CO2^{actual-}\right)}\right)^{0.5}$$

$$SD = \left(\frac{1}{N-1}\sum_{i=1}^{N}(error - error^{-})^{0.5}\right. \tag{12}$$

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N} error_i^2\right)^{0.5}$$

The MLR, GPR, and ANN models were constructed in the Mathematica 11.3 software using different input combinations of data (Table 3). As mentioned in the data description section, to predict CO_2 emission in the agriculture sector of Iran, we utilized EKC lecture that input variables as *GDP, GDP*^{2}*, Energy Consumption*, and *GINI* index.

3. Results and discussion

In first level MLR, GPR and ANN models were calibrated using the training data, and therefore, their training parameters are built-in. The best specification of each model was chosen according to the Taylor diagram. In the second level, an IMM model is used to treat the outputs of the best specification of the MLR, GPR, and ANN models as an ANN model input.

This section presents an overview of the results in Fig. 2, which show the Taylor diagrams of the models in the agriculture sector in Iran. In other previous studies, diagnostic criteria are expressed separately and there may be some contradiction between them. In this study, Table 4, presents the statistical analysis of the all examined models. The Taylor diagram provides a visual display by considering CC or R, SD, and RMSE at the same time for the modeled and observed (measured) CO_2 emission values. The closer the position of the modeled (predicted) values to that of the observed value, the better the model performance.

According to Fig. 2, among the MLR, GPR, and ANN models, the results for positions of the GPR4 (1.924), ANN4 (1.989), and MLR4 (2.046), respectively, are close to that of the observed value. It

Table 2
Information about parameters of ANN model in level 1 and 2.

Parameter	ANN model in level 1	ANN model in level 2
Input	Different combination of GDP, GDP^2, EC, GINI	Predicted CO_2 emission from best specifications of the MLR, GPR, ANN in level 1
Output	CO_2 emission	CO_2 emission
Number of hidden layers	2	2
Number of Neurons	5	5
Activation function in hidden layer	TanSigmoid	TanSigmoid
Activation function in output layer	TanSigmoid	TanSigmoid
Algorithm	Back Propagation	Back Propagation

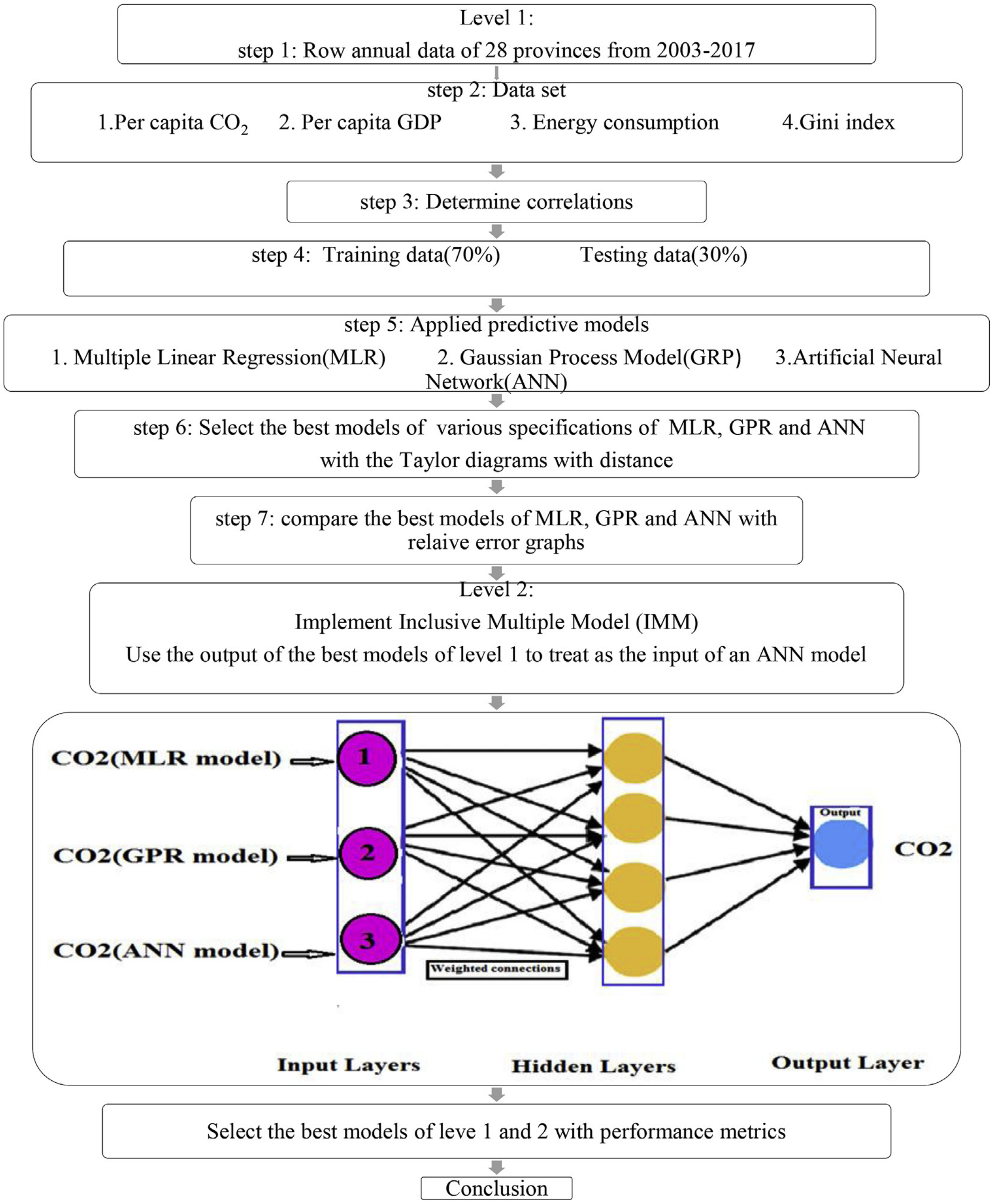

Fig. 1. The flowchart of the process of reviewing models and choosing the best model.

Table 3
Different input combinations of the MLR and GPR models.

Model	Input variables	Output variables
MLR1, GPR1, ANN1	GDP	CO_2
MLR2, GPR2, ANN2	GDP, GDP^2	CO_2
MLR3, GPR3, ANN3	GDP, GDP^2, EC	CO_2
MLR4, GPR4, ANN4	GDP, GDP^2, EC, GINI	CO_2
IMM	Predicted CO_2 with MLR4, GPR4, ANN4	CO_2

should be noted that all the models show almost a similar performance in estimating CO_2 emission; however, the GPR is more successful than MLR and ANN models. Besides, it can be seen that the predictive accuracy of the models increases each time by adding new explanatory variables to the model. The quadratic functional form has better performance than linear form as there is clearly a nonlinear and curve relationship between CO_2 emission and economic growth. The better result of the fourth specification in each model indicates that, in addition to per capita income, income distribution is also useful in examining the relationship between growth-pollution nexus, confirming the argument of Boyce (1994).

After implementing the IMM model, it was clarified that the performance of the IMM model is better than all the models in level 1. According to Fig. 3, the IMM model has the lowest distance from observation (1.857). In the other word, according to three performance metrics, the located position of the predicted values of CO_2 emission of the IMM model in space has the nearest distance with actual CO_2 emission values. These results suggest that the strategy developed leads to more accurate predictions with high correlation and lower error. So, the IMM model would predict CO_2 emission

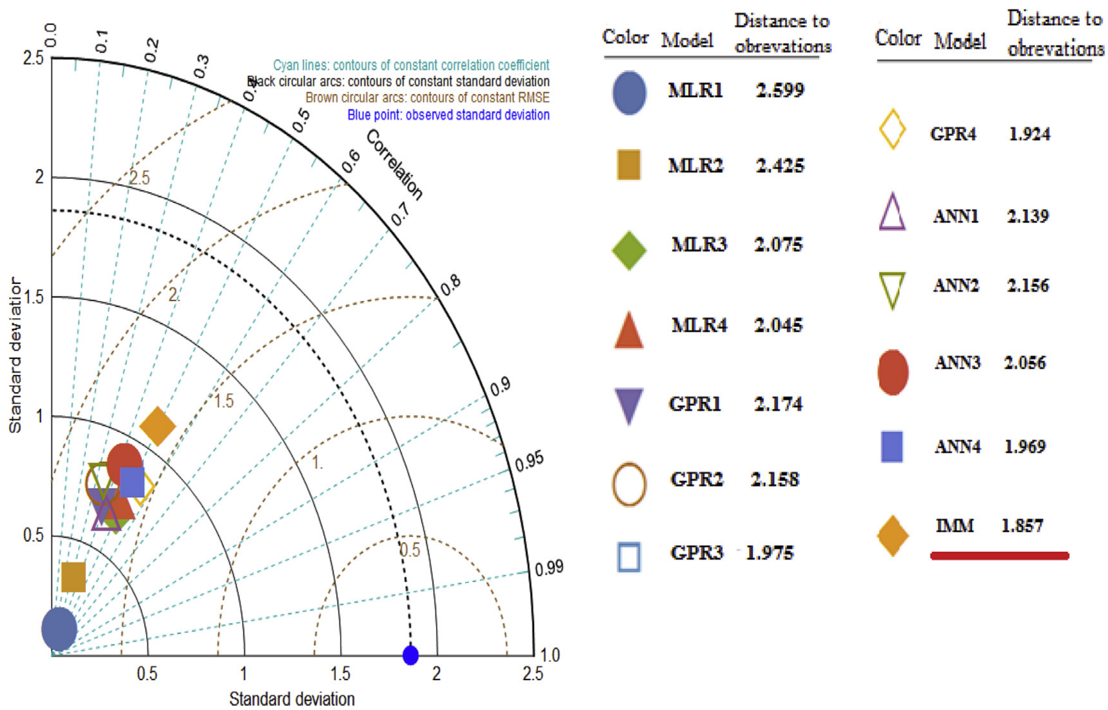

Fig. 2. The Taylor plots of the MLR, GPR and ANN models with different input combinations and the IMM model for prediction of CO_2 emission in the agriculture sector of Iran.

Table 4
Statistical analysis of the examined models.

Model	RMSE	R	SD	Model	RMSE	R	SD
MLR1	3.65	0.59	0.23	GPR4	0.89	0.78	0.45
MLR2	4.14	0.60	0.35	ANN1	3.85	0.62	0.73
MLR3	3.79	0.64	0.69	ANN2	3.47	0.66	0.45
MLR4	3.45	0.68	0.55	ANN3	1.50	0.70	0.89
GPR1	3.51	0.64	0.82	ANN4	1.32	0.76	0.56
GPR2	3.59	0.67	0.76	IMM	0.69	0.81	0.32
GPR3	1.34	0.73	0.88				

values more accurately.

In the next steps, the performance of the IMM model was compared with the best models of MLR, GPR, and ANN models by implementing various novel diagnostic graphs, including relative error graphs, a new version of scatter diagram with special colour spectrum for predicted and actual values, probability distribution function of residuals, and a new version of scatter diagram for residuals.

This section presents the relative error graphs for the models (Fig. 3). Each of these 140 points shown the difference between the predicted and actual CO_2 emission values in the testing phase. The relative errors show that what percentage of a forecasting model's errors are between −5 and 5. A model with a high percentage of the relative error will be the most accurate model to predict. The relative errors for MLR4, GPR4, ANN4, and IMM models are 24.12%, 31.42%, 25.28%, and 35.71%, respectively. The IMM model has the highest residuals between −5 and 5. So in the IMM model, predicted CO_2 values have the highest correlation with actual CO_2 values, and lowest errors, hence it can predict CO_2 emissions with low errors and high accuracy. Furthermore, the GPR4 model has high relative errors than the MLR4 and ANN4 models.

In this section, we plotted to scatter diagrams of the best models of level 1, and the IMM model in an innovatively and specially manner. The colour spectrum used helps us to understand better the relationship between predicted, and actual CO_2 emission values. The correlation coefficient (CC or R) and RMSE are two important measures to determine the accuracy of a model. A high CC and a low RMSE show that predicted and actual values in a model have a high correlation and so the accuracy of a model is higher.

Scatter diagrams of the MLR4, GPR4, ANN4, and IMM models (Fig. 4) indicate that the IMM model has the highest correlation coefficient and the lowest RMSE among the other models, followed by the GPR4, ANN4, and MLR4 models. These metrics show that the IMM model has higher accuracy than the other examined models. These results are in line with the results of studies showing the superiority and accuracy of new predicting methods over traditional methods such as regression and neural network. Besides results indicated that according to these metrics, the GPR4 model had better performance than MLR4 and ANN4 models.

Another graph that can help to compare the performance of the examined models to predict CO_2 emission values is the residuals probability distribution function (PDF). The PDF of the residuals determines the mean, standard deviation, and frequency of each range of the errors. The PDF of the models is presented in Fig. 5. The model with residuals PDF closer to the normal distribution (with mean = 0 and standard deviation = 1) is better in prediction; the tip of the curve is also sharper and fewer errors are more frequent.

From Fig. 5, it is apparent that the residuals PDF of the IMM model is closer to the normal distribution (mean = 0.058 and standard deviation = 1.540) with the highest frequency of fewer errors, the frequency of the errors between −5 and 5 is 35%. For the MLR4, GPR4, and ANN4 models the frequency of the errors between −5 and 5 is 24%, 30%, and 26%, respectively. Beside the GPR4 model has the nearest residuals PDF to normal distribution after the IMM model. So, according to residuals PDF, the predicted values of the IMM model are more accurate.

We have also plotted an innovation scatter with a unique colour spectrum for the residuals of MLR4, GPR4, ANN4, and IMM models. This colour spectrum helps to determine better the situation of the

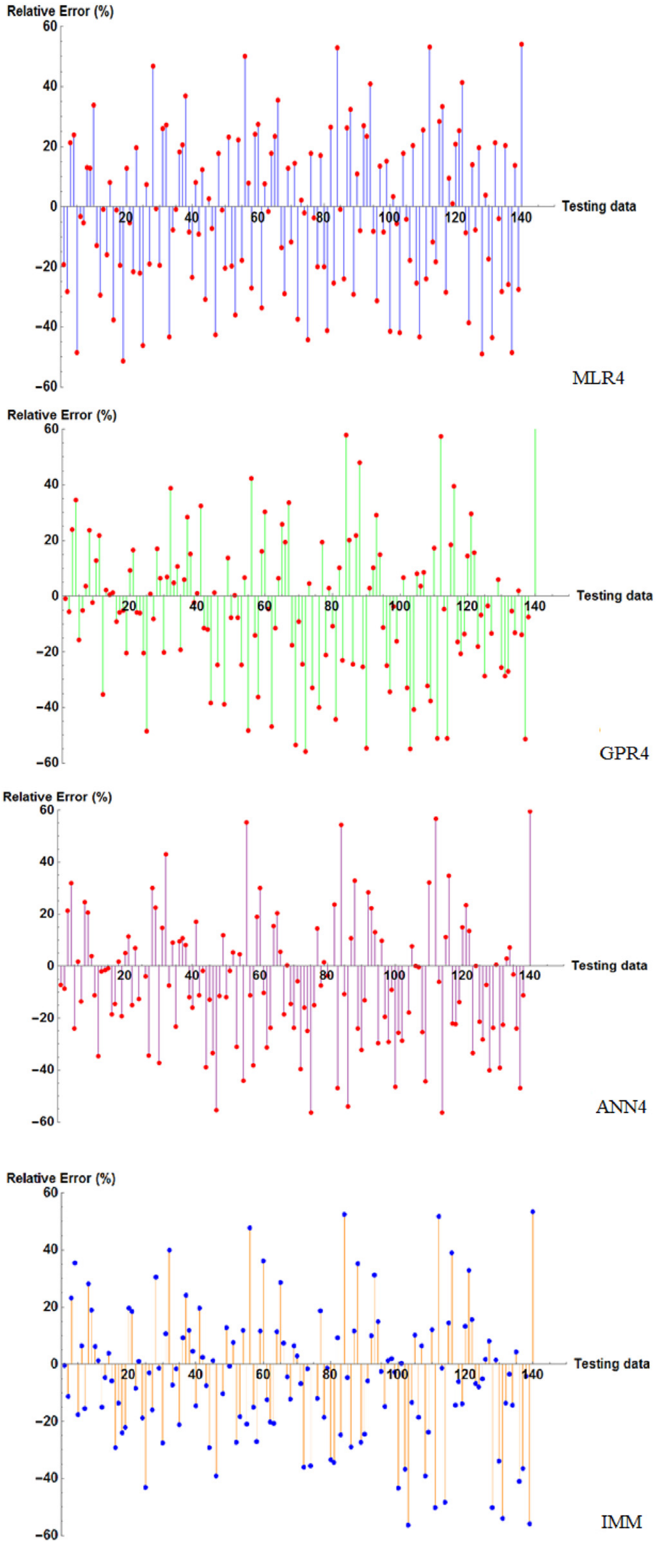

Fig. 3. Relative error graphs for MLR4, GPR4, ANN4, and IMM models.

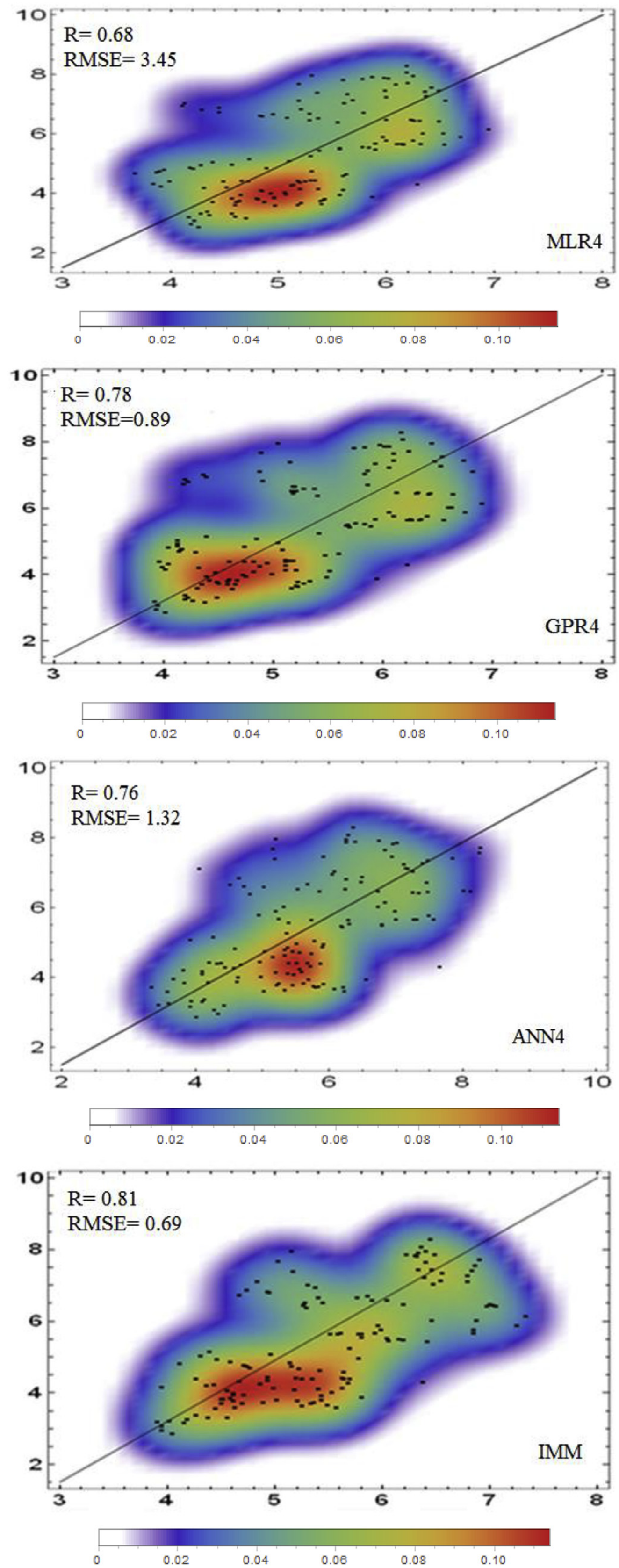

Fig. 4. Scatter diagrams for the MLR4, GPR4, ANN4, and IMM models.

errors of the predicted values of the models, and making the right decision about choosing a more accurate model with smaller errors. Fig. 6 represents the scatter of residual for MLR4, GPR4, ANN4, and IMM models. It is seen that errors are more concentrated around 0 for the IMM model than the other three models indicating that a large number of data points display a good correlation and

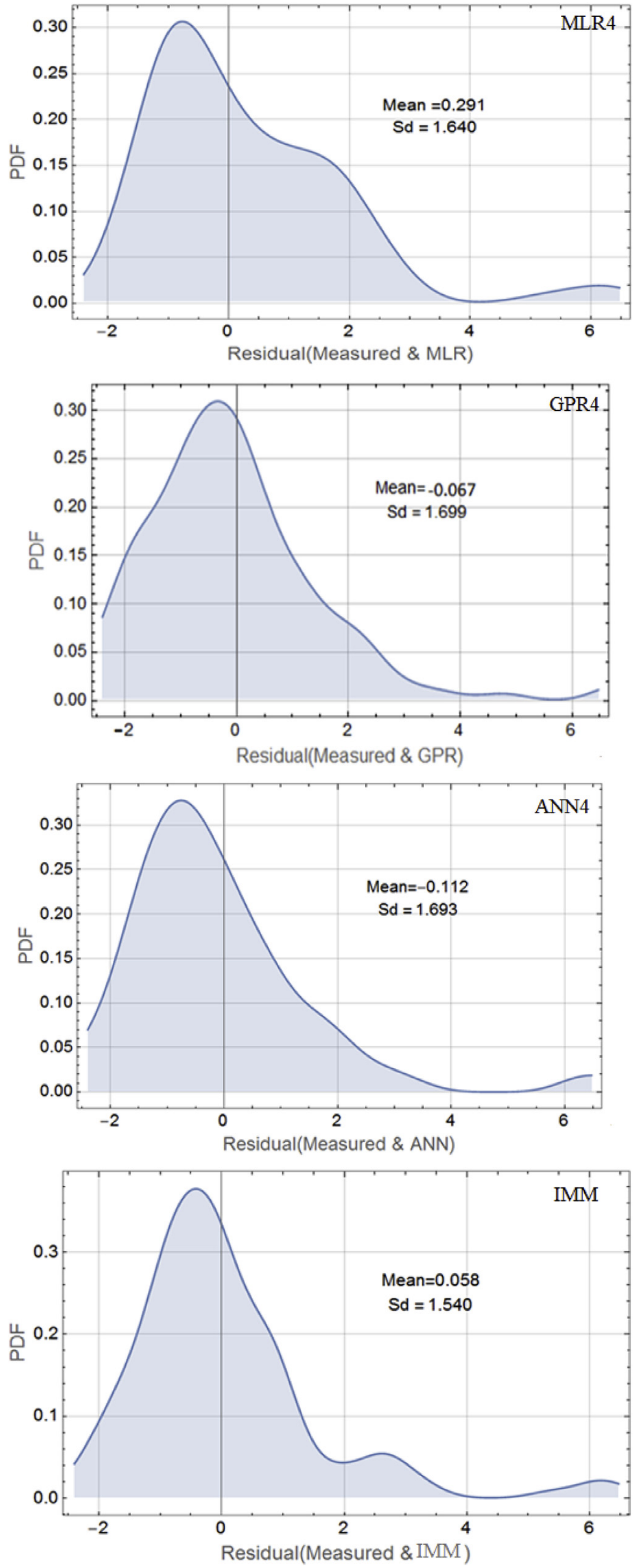

Fig. 5. PDF plot of residuals for MLR4, GPR4, ANN4, and IMM models.

discordance is not very strong at higher ranges, suggesting the high accuracy of the IMM model.

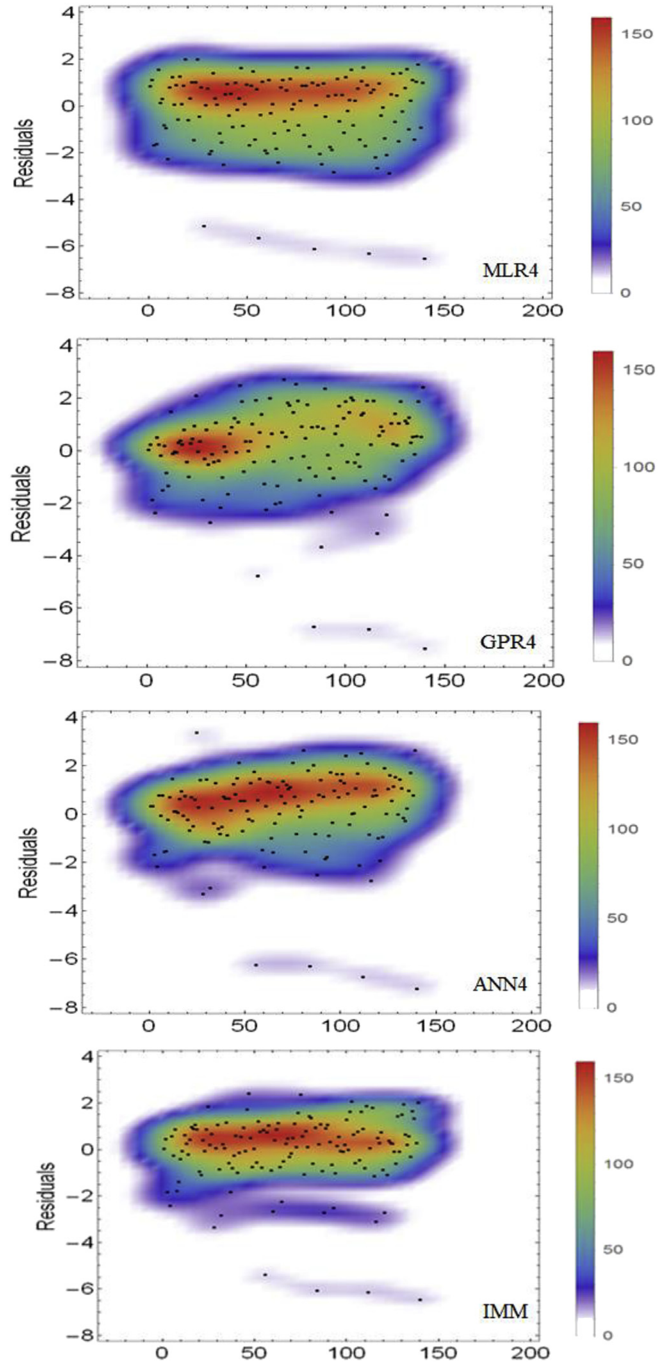

Fig. 6. Residual plots for MLR4, GPR4, ANN4, and IMM models.

Due to the high sensitivity of the carbon emission issue, and the high importance of prediction as a valuable policy tools to control it, one of the main challenges in policymaking to control and reduce carbon emission have been developed a model with high-precision predictive power. Determining the most accurate model is another issue for researchers, especially when the number of the examined models in a study is high or the criteria are inconsistent. However, providing a single criterion that can consider several evaluation criteria simultaneously can solve this problem and minimize decision errors. So, developing a novel high accurate approach to predict CO_2 emission, and presenting unique criteria named distance to observation point were the most aims of this paper.

Despite the considerable advantages and high predictive power of the IMM model, it has not been implemented in environmental economics and pollution emission prediction. This is while the results of this paper indicated the high accuracy of the IMM in predicting CO_2 emission. Because according to various criteria and our innovation diagrams, it is observed that the predicted CO_2 emission values with the IMM model have less distance to the observed CO_2 emission values. The IMM model had high prediction accuracy, with having a higher frequency of the relative errors, the highest correlation coefficient, and less RMSE than the MLR4, GPR4, and ANN4 models, that these make the results of the IMM model more reliable, and it assures us that the strategy developed improves the results.

These findings are similar to those of Khatibi et al. (2017) and Kashani et al. (2020), who indicated that the IMM model to be a more accurate approach than other models used in their studies, such as the Firefly algorithm, M5Tree, MARS, ANN, and SVM models. The IMM model, therefore, could be a precise model for predicting carbon emission trends in the future. Thus, planners, policymakers, and institutions, who are concerned about the devastating effects of air pollution, could implement the model to adopt appropriate programs such as taxation and incentives to reduce carbon emissions. Due to the high accuracy and advantages of the IMM model, the use of this model in modeling and predicting in economics, environmental issues, and other scientific fields are recommended. Besides, the graphs introduced in our study can be useful for making the correct decisions and reducing errors in determining the most accurate prediction model.

4. Conclusion

In the present paper, we developed the IMM as a novel approach to predict CO_2 emission in the agriculture sector of Iran, as well as implemented novel colorful scatter diagrams, and a criteria called distance to the observation point to determine the most accurate model. The emergence of the present study is in terms of the importance of accurate prediction for planning to control carbon emission as a serious threat to human health, climate change, and global warming.

Results indicated that there is a nonlinear relationship between economic growth and CO_2 emission, and income inequality should be considered as a primary factor in EKC evaluation. The lowest distance from the observation point (1.857), the highest correlation coefficient (0.81), the lowest standard deviation, the lowest RMSE (0.69), the highest errors between -5 and 5 (37.84%), and the closest residual's PDF to the normal distribution confirmed the high accuracy of our developed model compared to other examined models.

Governments, policymakers, local, and international institutes, who are concerned about hazardous effects of air pollution, are recommended to implement the IMM as the most accurate model to predict CO_2 emission for planning to reduce its concentration. Distance to the observation point could be the perfect criteria to determine the most accurate model without personal decision errors. The results indicated that there are possible improvements in approaches to predict pollutants and also towards the next-generation research activities.

CRediT authorship contribution statement

Elham Shabani: Conceptualization, Methodology, Software, Writing - original draft. **Babollah Hayati:** Project administration, Supervision, Data curation. **Esmaeil Pishbahar:** Methodology, Validation. **Mohammad Ali Ghorbani:** Software, Formal analysis, Writing - review & editing. **Mohammad Ghahremanzadeh:** Investigation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdelfatah, A., Mokhtar, S.A., Sheta, A., 2013. Forecast global carbon dioxide emission using Swarm intelligence. Int. J. Comput. Appl. 77 (12), 1–5.

Ann, J., Shin, D., Kim, K., Yang, J., 2017. Indoor air quality analysis using deep learning with sensor data. Sensors 17 (11), 2476. https://doi.org/10.3390/s17112476, 2017.

Alam, M.M., Murad, M.W., Noman, A.H.M., Ozturk, I., 2016. Relationship among carbon emission, economic growth, energy consumption and population growth: testing environmental Kuznets curve hypothesis for Brazil, China, India and Indonesia. Ecol. Indicat. 70, 466–479.

Apergis, N., Payne, J., 2009. CO_2 emissions, energy usage and output in Central America. Energy Pol. 37, 3282–3286.

Aydin, C., Esen, O., 2018. Reducing CO_2 emissions in the EU member states: do environmental taxes work? J. Environ. Plann. Manag. https://doi.org/10.1080/09640568.2017.139573.

Behrang, M.A., Assareh, E., Assari, M.R., Ghanbarzadeh, A., 2011. Using Bees algorithm and artificial neural network to forecast world carbon dioxide emission. Energy Sources, Part A Recovery, Util. Environ. Eff. 33 (19), 1747–1759.

Ben Jebli, M., Yussef, S.B., Ozturk, I., 2016. Testing environmental Kuznets curve hypothesis: the role of renewable and non-renewble energy consumption and trade in OECD countries. Ecol. Indicat. 60, 824–831.

Borghesi, S., 2000. Income inequality and the environmental Kuznets curve. NOTA DI LAVORO 83, 2000.

Boyce, J.K., 1994. Inequality as a cause of environmental degradation. Ecol. Econ. 11 (3), 169–178.

Chen, Sh, Mihara, K., Wen, J., 2018. Time series prediction of CO_2, TVOC and HCHO based on machine learning at different sampling points. Build. Environ. 146, 238–246.

Erol, U., Yu, E.S.h., 1987. Time-series analysis of the causal relationships between US energy and employment. Resour. Energy 9, 75–89.

Fang, D., Zhang, X., Yu, Q., Jin, T.Ch, Tian, L., 2018. A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression. J. Clean. Prod. 173, 143–150.

Global carbon atlas, 2018. available at www.globalcarbonatlas.org.

Galeotti, M., Manera, M., Lanza, A., 2009. On the robustness of robustness checks of the environmental Kuznets curve hypothesis. Environ. Resour. Econ. 42 (4), 551–574.

Ghorbani, M.A., Khatibi, R., Karimi, V., Zaher Mundher, Y., Zounemat-Kermani, M., 2018. Learning from multiple models using artificial intelligence to improve model prediction accuracies: application to river flows. Water Resour. Manag. https://doi.org/10.1007/s11269-018-2038-x.

Girard, A., Rasmussen, C.E., Quiñonero-Candela, J., Murray-Smith, R., 2003. Gaussian Process priors with uncertain inputs_ application to multiple-step ahead time series forecasting. In: Neural Information Processing Systems, Vancouver, Canada, ISBN 0262025507.

Golley, J., Meng, X., 2012. Income inequality and carbon dioxide emissions: the case of Chinese urban households. Energy Econ. 34, 1864–1872.

Hao, Y., Chen, H., Zhang, Q., 2016. Will income inequality affect environmental quality? Analysis based on China's provincial panel data. Ecol. Indicat. 67, 533–542.

He, Y., Lin, B., 2019. Investigation environmental Kuznets curve from an energy intensity perspective: empirical evidence from China. J. Clean. Prod. 234, 1013–1024.

Hu, J., Wang, J., 2015. Short-term wind speed prediction using empirical wavelet transform and Gaussian process regression. Energy 93, 1456–1466.

Jalil, A., Mahmud, S.F., 2009. Environmental Kuznets curve for CO_2 emission: a cointegration analysis for China. Energy Pol. 37, 5167–5172.

Kardani, M.N., Baghban, A., Sasanipour, J., Mohammadi, A.H., Habibzadeh, S., 2018. Group contribution methods for estimating CO_2 absorption capacities of imidazolium and ammonium-based polyionic liquids. J. Clean. Prod. 203, 601–618.

Kashani, M., Ghorbani, M.A., Shahabi, M., Raghavendra, N., Diop, L., 2020. Multiple AI model integration strategy—Application to saturated hydraulic conductivity prediction from easily available soil properties. Soil and Tillage Res. 196. https://doi.org/10.1016/j.still.2019.104449.

Kasman, A., Duman, Y., 2015. CO_2 emission, economic growth, energy consumption, trade and urbanization in new member and candidate countries: a panel data analysis. Economic Modeling 44, 97–103.

Khatibi, R., Ghorbani, M.A., Akhuni Pourhosseini, F., 2017. Stream flow predictions using nature-inspired Firefly Algorithms and a Multiple Model strategy – directions of innovation towards next-generation practices. Adv. Eng. Inf. 34, 80–89.

Khatibi, R., Ghorbani, M.A., Naghshara, S., Aydin, H., Karimi, V., 2020. A framework for 'inclusive multiple modelling' with critical views on modelling practices – applications to modelling water levels of caspian sea and lakes urmia and van. J. Hydrol. 287, 124923, 2020.

Lee, C.-C., Lee, J.D., 2009. Income and CO_2 emissions: evidence from panel unit root

and cointegration tests. Energy Pol. 37, 413–423.
Libo, Y., Tingting, Y., Jielian, Zh., Guicai, L., Yanfen, L., Xiaoqian, M., 2017. Prediction of CO_2 emissions based on multiple linear regression analysis. Energy Procedia 105, 4222–4228.
Lin, ChS., Liou, F.M., Huang, ChP., 2011. Grey forecasting model for CO_2 emissions: a Taiwan study. Appl. Energy 88 (11), 3816–3820.
Lotfalipour, M.R., Falahi, M.A., Bastam, M., 2013. Prediction of CO_2 emissions in Iran using grey and ARIMA models. Int. J. Energy Econ. Pol. 3 (3), 229–237.
Magnani, E., 2000. The environmental Kuznets curve, environmental protection policy and income distribution. Ecol. Econ. 32 (3), 431–443.
Noori, R., Moazami, S., Jabbarian Amiri, B., Yeganeh, B., Partani, S., Safavi, S., 2015. Reliable prediction of carbon monoxide using developed support vector machine. Atmospheric Pollution Research 1–7.
Oliveira, V., Sousa, V., Dias-Ferreira, C., 2018. Artificial neural network modelling of the amount of separately-collected household packaging waste. J. Clean. Prod. https://doi.org/10.1016/j.jclepro.2018.11.063.
Ozcan, B., 2013. The nexus between carbon emissions, energy consumption and economic growth in Middle East countries: a panel data analysis. Energy Pol. 62, 1138–1147.
Pao, H.T., Fu, H.Ch, Tseng, ChL., 2012. Forecasting of CO2 emissions, energy consumption and economic growth in China using an improved grey model. Energy 40, 400–409.
Pao, H.T., Tsai, Ch M., 2011. Modeling and forecasting the CO_2 emissions, energy consumption, and economic growth in Brazil. Energy Pol. 35, 58–69.
Pata, U.K., 2018. Renewable energy consumption, urbanization, financial development, income and CO_2 emissions in Turkey: testing EKC hypothesis with structural breaks. J. Clean. Prod. 187, 770–779.
Pearson, K., Lee, A., et al., 1908. On the generalized probable error in multiple normal correlation. Biometrica 6, 59–68.
Quiñonero-Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. 6, 1939–1959.
Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press.
Ridzuan, S., 2019. Inequality and the environmental Kuznets curve. J. Clean. Prod. https://doi.org/10.1016/j.jclepro.2019.04.284.
Saleh, Ch., Dzakiyullah, N.R., Bayu, J., 2015. Carbon dioxide emission prediction using support vector machine. In: Materials Science and Engineering Conference.
Samsami, R., 2013. Application of Ant Colony optimization (ACO) to forecast CO_2 emission in Iran. Bulletin of Environment, Pharmacology and Life Sciences2 (6), 95–99.
Sangeetha, A., Amudha, T., et al., 2018. A novel bio-inspired framework for CO_2 emission forecast in India. Procedia Comput. Sci. 125, 367–375.
Sun, W., Liu, M., 2016. Prediction and analysis of the three major industries and residential consumption CO_2 emissions based on the least squares support vector machine in China. J. Clean. Prod. 122, 144–153.
Taghavifar, H., Taghavifar, H., Mardani, A., Mohebbi, A., Khalilarya, Sh, Jafarmadar, S., 2015. Appraisal of artificial neural networks to the emission analysis and prediction of CO_2, soot, and NOx of n-heptane fueled engine. J. Clean. Prod. 1–11.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysics Resource Atmospheric 106, 7183–7192.
Uzar, U., Eyuboglu, K., 2019. The nexus between income inequality and CO_2 emission in Turkey. J. Clean. Prod. 227, 149–157.
Wang, Y.N., Kang, Y.Q., Wang, J., Xu, L.N., 2017. Panel estimation for the impacts of population-related factors on CO_2 emissions: a regional analysis in China. Ecol. Indicat. 78, 322–330.
Wang, ZhX., Li, Q., 2019. Modeling the nonlinear relationship between CO_2 emissions and economic growth using a PSO algorithm-based grey Verhulst model. J. Clean. Prod. 207, 214–224.
Xu, G., Schwarz, P., Yang, H., et al., 2019. Determining China's CO_2 emissions peak with a dynamic nonlinear artificial neural network approach and scenario analysis. Energy Pol. 128.
Yu, Y., Deng, Y.R., Chen, F.F., 2017. Impact of population aging and industrial structure on CO_2 emissions and emissions trend prediction in China. Atmospheric Pollution Research. https://doi.org/10.1016/j.apr.2017.11.008.
Zarei, F., Rahimi, M.R., Razavi, R., Baghban, A., 2019. Insight into the experimental and modeling study of process intensification for post-combustion CO_2 capture by rotating packed bed. J. Clean. Prod. 211, 953–961.
Zhang, L., Shen, Q., Wang, M., Sun, N., Wei, W., Lei, Y., Wang, Y., 2019. Driving factors and predictions of CO_2 emission in China's coal chemical industry. J. Clean. Prod. 210, 1131–1140.