Hindawi
Complexity
Volume 2022, Article ID 2631939, 20 pages
https://doi.org/10.1155/2022/2631939

WILEY | Hindawi

Research Article

Forecasting Methods in Various Applications Using Algorithm of Estimation Regression Models and Converting Data Sets into Markov Model

Mohammed M. El Genidy[1] **and Mokhtar S. Beheary**[2]

[1]*Department of Mathematics and Computer Science, Faculty of Science, Port Said University, Port Said, Egypt*
[2]*Department of Environmental Science, Faculty of Science, Port Said University, Port Said, Egypt*

Correspondence should be addressed to Mohammed M. El Genidy; drmmg2016@yahoo.com

Received 30 September 2021; Revised 18 December 2021; Accepted 28 December 2021; Published 29 January 2022

Academic Editor: Qingling Wang

Water quality control helps in the estimation of water bodies and detects the span of pollutants and their effect on the neighboring environment. This is why the water quality of the northern part of Lake Manzala has been studied here from January to March, 2016. This study aims to model and create a program for linear and nonlinear regression of the water elements in Lake Manzala to assess and predict the water quality. Water samples have been extracted from various depths, and physio-chemical properties and heavy metal concentrations have been evaluated. This study has proposed a new algorithm for predicting water quality called "Algorithm of Estimation Regression Model" (AERM). On the contrary, in renewable energy applications, statistical modeling and forecasting the solar radiation remains a significant issue with detect to reinforce power management. A new proposed method for forecasting the average Monthly Global Solar Energy (MGSE) in Queensland, Australia, is called, Converting Data Set into Markov Model (CDMM). It was used to obtain Markov transition probability matrices for three and six states of the solar energy. The proposed forecasting method yielded accurate results with minimal error.

1. Introduction

Adequate amounts of suitable-quality water resources provide a precondition for economic development and ecological integrity. Numerous stresses influence water quality, such as natural processes (e.g., weathering, precipitation, and soil erosion), anthropogenic activities (e.g., agricultural, urban, and industrial activities), and the increased utilization of water resources. Because of multifaceted effects noted above, water quality deterioration has become a serious issue worldwide as Wu et al. and Qin et al. emphasized [1, 2].

In recent years, Elkady et al. showed that Lake Manzala could have been recorded as the largest and the most productive Lake of the northern Egyptian coastal lakes [3]. It is a shallow, brackish-water lake located on the northeastern edge of the Nile Delta between Port Said and the Suez Canal eastwards and the Damietta brink of the Nile westwards. It is an important natural resource for fish catch, wildlife, hydrologic and biologic regimes, and table salt production. Lake Manzala is characterized by special sensitive environments.

The lake has declined in area over the last two decades, reaching 700 km^2 in 2003, as mentioned by Ali [4]. The lake has five major drains which carry agricultural drainage water as well as waste water from urban areas. It crosses the Delta and discharges into the southern, eastern, and western parts of the Lake Manzala, as discussed by Abukila, Khadr, and Elshemy [5, 6].

It is also essential to enhance water quality because it plays a pivotal role in ecological and human health and in economic development. Based on information from assessments, the public is more likely to implement a protective measure that improves the conditions of water bodies, studied by Jiang et al. and Sousa et al. [7, 8].

In this study, estimation of water quality using regression models has been applied to assess the water quality and its spatial variations in Lake Manzala. The primary objectives of this study are to determine the water quality status, whereby its spatial variation across the study area and to explore the critical parameters in the development of a statistical model for simple and cost-effective water quality evaluation. The performance of the statistical model will improve with full consideration of parameter estimation.

Several researchers have developed monthly mean global solar radiation models that include classical empirical regression, artificial neural networks (ANNs), and time-series regression techniques, auto regression moving average (ARMA), and empirical correlation to estimate the monthly mean daily global solar radiation (GSR). Researchers have faced serious difficulties in modeling, estimating, and obtaining the prediction equation of meteorological data like the solar energy amount and the corresponding maximum temperature. A variety of methods for estimating global solar exposure have been published using empirical correlations. A model was proposed depending on the statistic properties of hourly global radiation with a variety of parameters. It represents an accurate understanding of hourly solar radiation, performed by Aguiar and Collares-Pereira [9].

Almost all approaches used applied forecasting parameters such as period of sunlight, temperature degree, and moisture. Methods must therefore include more than two parameters to obtain reliable tests, which can carry more complex and higher computational error risks. Relying on their ability to make statistical decisions, they have noticed that the sampling method turns into statistical inference methods. Thus, they need to select the appropriate statistical inference method for the terminal objective of the significant study. Here lies the importance of the estimation method.

The development of pattern similarity in solar radiation estimation is presented in the clustering algorithm and its application. For the extraction of shape-based clusters from the input meteorological parameters, the continuous-density hidden Markov model (HMM) with Pearson's R model is used and then processed by the generalized fuzzy model (GFM) to accurately estimate solar radiation. The patterns of the data vectors are used as the similarity index for clustering instead of using distance function as an index of similarity. Here, which overcomes few of the drawbacks associated with distance-based approaches to clustering.

The method of estimation used here exploits the HMM's pattern recognition prowess for the collection and generalization of clusters and the GFM's nonlinear modeling capabilities in forecast the solar radiation. The proposed model is applied to 15 different classifications of various combinations of meteorological parameters, applied by Bhardwaj et al. [10]. As standard practice is to use acceptable empirical correlations to estimate the average daily global solar radiation based on the applicable data measured at those locations. For few regions, there are no real calculated values, and these correlations estimate the values of meteorological data for a region of investigation from more widely accessible meteorological, climatology, and geographical parameters.

Some researchers show an interest to collect and study the comprehensive global models of solar radiation available in the literature chronologically. In addition, to identify them into four groups, such as sunshine-based, cloud-based, temperature-based, as well as other meteorological parameter-based models, depending on the meteorological parameters used as model data, showed by Besharat et al. [11]. The artificial neural network (ANN) is used to identify effective methods available for solar radiation prediction in the literature and to identify research gaps. In contrast to traditional approaches, artificial neural network techniques forecast solar radiation more accurately. The prediction accuracy of an ANN model is found to be based on the combination of input parameters, training algorithms, and configuration of architecture, carried out by Yadav and Chandel [12].

In particular, artificial neural network was performed to predict the direction movement of financial time series. The resilient back-propagation learning procedure was suggested for training a single-layer feed-forward neural netwo. The data set of the network was 15 indicators, and the single output of the daily stock closing price takes the value of either 0 or 1. Whereby, 0 indicates that the daily closing stock price index for the next day is lower than today's price and 1 means that it is higher. If the output value ≥ 0.5, the prediction of direction movement is considered upward, and if it is less than 0.5, it is considered downward [13].

It is clear from this previous study that it was based only on two states in the output, namely, zero and one, in which the stock exchange is closed, and the prediction will be on the next day of closing the stock exchange. In addition, this method needs to improve the parameter values of neural networks, which leads to the creation of complex algorithms. On the contrary, this study was to obtain a prediction using Markov models for three states of solar radiation in addition other six states for each month during the successive years and do not only predict the next state for once, nor only two states "0" and "1" like in the previous study.

In addition, it is possible to create a program that is not as complicated as the one in the previous study, and the accuracy can be increased by increasing the number of states, and it does not need to improve any parameters or indicators, as is the case in the previous study.

On the contrary, the prediction for various drought classes using spatiotemporal categorical sequences relied on the forecasting method on drought data for the interval from 1971 to 2017, which were divided into six cases while they used only one optimal probability distribution named three-parameter Weibull distribution for all drought data during (1971–2017) in a specific region where there is a monitoring station named Astore station with time scale equal one. In addition, the logistic regression model was considered to compute the probability of drought persistence from one season to the next one to survey the seasonal drought frequency and drought persistence in the northern area of Pakistan [14, 15].

Therefore, the aforementioned studies relied on the prediction method during the months of the same year, with a single probability distribution for all data. While the droughts during the different months are considered independent, which leads to prediction error and inaccuracy, they did not calculate the estimation error in the forecast. As they did not show that the sum of the state probabilities in each column in the transition matrix equal one using three-parameter Weibull, it is known that each column in the transition matrix will be represented by a single distribution due to the different transition states for each column. For this reason, the paper should study the drought classes for each month separately during consecutive years. Each month should be represented by a transition matrix which represents the transition states during successive years to predict the state of each month more accurately.

It is remarkable that the effects of component parameters from the mixing stage on the manufactured results of Li-ion battery electrode via classification modeling has been studied. As suggested, an effective RUBoost-based ensemble learning framework to recover for category imbalance case and good classification of three key quality indicators for both LiFePO4- and Li4Ti5O12-based electrodeIn another previous study, a random forest (RF)-based classification framework, using the out-of-bag (OOB) predictions for effective quantification the importance of battery manufacturing advantages [16, 17].

While the prediction in this study was related to the field of natural resources in water quality by estimating its components in the different depths of the lakes, which have not been measured before, the sample is limited and has a high dispersion. There is always a need for a large sample size of water element measurements to be represented graphically by functions of probability distributions containing three parameters: shape, scale, and location. Thus, the probabilities of the outliers of the elements can be predicted, as well as the skewness and kurtosis and the measurement range of each element in the water at different depths. The application of the AERM algorithm is very possible in the manufacture of batteries and that is on the measurement points of each of the lifetime and corresponding battery voltage. Therefore, the battery voltage values can be represented by a cumulative probability distribution function that contains three parameters, and it is of great importance in increasing the lifetime and reliability and reducing the failure rate of the battery.

It was also recorded that an algorithm is called the advanced proximal policy optimization (PPO) reinforcement learning was used to improve speed control of the model-free quadrotor. It was used for acquired neural networks to determine the states of the command control system in an end-to-end style. In addition, it has been suggested as an adaptive neural network (NN) distributed control algorithm for a group of high-order nonlinear agents with nonidentical unknown control directions (UCDs) under signed time-varying topologies [18, 19].

In this study, AERM's algorithm determined the unknown readings at various other depths of the lake, which were not measured by water element analyzers. In addition, the sample is limited and has high dispersion. AERM's algorithm accurately estimates the unknown readings by controlling the values of R-square (R^2), the total sum of squares (SST), the regression sum of squares (SSR), and the error sum of squares (SSE). Nonetheless, AERM algorithm can be applied on four acquired neural networks to obtain a larger number of estimations of other unknown states of the command control system. For this reason, AERM algorithm may reduce time, cost, and effort in obtaining other unknown states or measurements of the command control system. In contrast with the method of CDMM dealing with a lot of size measurements, the aim was to convert these measurements to a limited number of states. Thus, the idea of CDMM method can also be used on the control systems, which have a large number of different measurements.

2. Complex Forecasting Science

2.1. Why Complex Forecasting Science? Complex forecasting science considered on sampling and analysis of data sets for different applications such as water quality and solar energy. In this study, a new estimation method called algorithm of estimation regression model (AERM) was implemented to evaluate the water elements in the lake. This algorithm was characterized by its high accuracy more than the estimation methods found in the previous studies.

The eastern part of Lake Manzala is separated from the Mediterranean Sea by a sandy beach ridge and is isolated from the main lake by a coastal road. The study area receives its feed from the Mediterranean Sea through El-Gamil and El Mussallas inlets and accepts drainage water from Ezbet El Borg drain. Extracted water samples from various depths in Lake Manzala in Egypt, physio-chemical properties, and heavy metal concentrations have been evaluated, and all sampling sites are displayed in Figure 1, which presents 12 different sampling sites of Lake Manzala in Egypt for water quality, studied by Beheary and El-Matary [20].

This study has been based on a data set of 18 parameters, measured three times from January to March 2016, at 12 sampling sites that cover the northern part of Lake Manzala at the fish farming area called El Mussallas (see Figure 1). Sampling locations have been identified using global positioning system (GPS), while water samples have been analyzed in the field using multiparameter water quality probe (Aqua Prob. AP-2000) for Depth (m), Temp (C), PH, ORP (REDOX), DO (mg/L), EC (uS/cm @25°C), RES (Ohms.cm), TDS (mg/L), SAL (PSU), and SSG (st). According to APHA, water samples were brought to the laboratory and analyzed for TN (mg/L), COD (mg/L), and total phosphorous (mg/L).

UNEP/IAEA showed that the total heavy metal in water was measured as follows. An exact 100 mL of the sample is to be placed into a beaker, then 5 mL concentrated HNO_3 is to be added. That is to be carefully boiled on a hot plate or a steam bath until it evaporates down to about 20 mL. Then, a further 5 mL concentrated HNO_3 is to be added and covered with a watch glass to be heated this way. The process of adding and heating concentrated HNO_3 is to continue until the solution appears colored light and clear. This sign indicates that digestion is complete, yet drying is

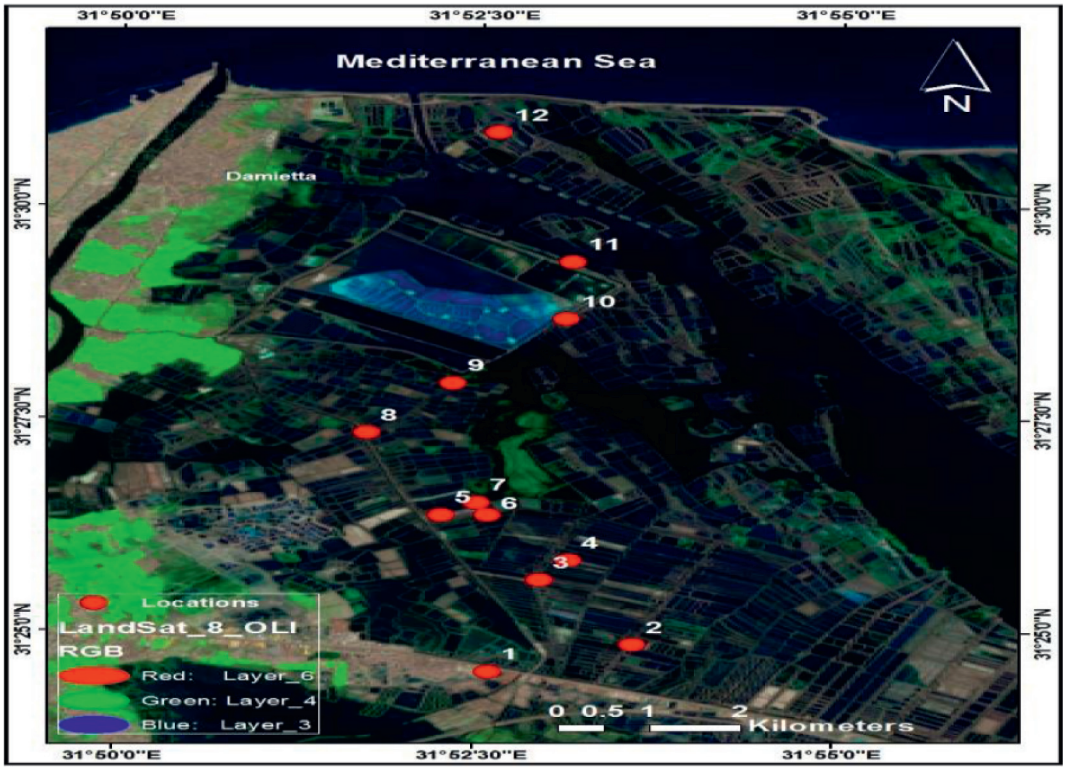

Figure 1: Sampling sites from Lake Manzala, Egypt.

not allowed during the digestion. About 1to 2 mL concentrated HNO_3 is to be added and slightly heated to dissolve any remaining residue. The beaker walls are to be carefully washed down. Then, the entire content is to be poured into a 50 mL volumetric flask to cool. This is how all digested solutions have been analyzed for heavy metal (Cu, Cd, Pb, Zn, and Fe) using an atomic absorption spectrophotometer (Shimadzu AA-6800), and the results are canonically expressed as mg/L.

Water samples have been extracted from various depths of the northern part of Lake Manzala. Then, their physiochemical properties and heavy metal concentrations of Cu, Cd, Pb, Zn, and Fe have been estimated. Their respective mean estimations are sorted out in a tabular form to show how they differ at various depths ranging from 0.13 to 2.5 m. The values of these estimates are presented in Table 1, where (1) refers to the actual values of the elements in Lake Manzala, pH: potential of hydrogen, ORP: oxidation reduction potential, DO: dissolved oxygen, EC: electrical conductivity, RES: RESISTIVITY, TDS: total dissolved solids, SAL: SALINITY, SSG: seawater specific gravity, TN: total nitrogen, COD: chemical oxygen demand, PO4: total phosphorous, cd: cadmium, cu: copper, zn: zinc, pb: lead, fe: iron, Temp: temperature, C: degree Celsius, mg/L: milligrams per litre.

In this section, the statistical regression equation of the water quality has been performed to evaluate the water quality and its spatial variations in Lake Manzala. The prediction programs are a promising solution that could aptly be used to control the water quality of lakes insofar as real estimations of some specific locations.

In Queensland (Terrey Hills)-Australia, the annual solar energy map shows a higher energy of the south coastal regions to central and northern Australia with a lower solar energy. Inland Australia areas have a lower humidity in the air and thus less cloud cover. The data set used in this study included yearly global solar energy average (YGSEA). Climate data over 30 years in Queensland, Australia. It was actual and reliable data, as they are recorded by the Bureau of Meteorology in Australian government. Figure 2 indicates yearly global solar energy average (GSEA) in 30 years successively.

On the contrary, the largest amount of solar energy in Queensland, Australia, is distributed over January, February, September, October, November, and December. The month with the largest amount of solar energy. January and June are the months with the lowest amount of solar energy as indicated in Figure 3. The percentages of each month over 30 years are shown in Figure 4.

Typical values for daily global solar exposure range from 1 to 35 MJ/m^2 (mega joules per square meter) and are usually highest in clear sky conditions during the summer and lowest during the winter or very cloudy days.

2.2. What Is Complexity? This study attempts to estimate the values of water elements at various depths during significant time span. This is performed with the least unavoidable error rate at different values of depth.

The depth points were divided into two or more points, and the regression model was derived for each interval of water depth depending on the error sum of squares (SSE)

Table 1: Actual values of water elements at different depths of lake.

Depth (m)	0.13	0.29	0.3	0.31	0.33	0.38	1.34	1.5	2	2.5
Temp1 (°C)	22.1	22.2	22.833	22.6	21.8	22.3	19.9	20.7	20.2	20.7
pH1	8.54	8.51	7.153	8.69	9.09	8.37	8.42	7.92	8.11	8.72
ORP1 (REDOX)	−50.5	49.7	171.7	40.5	8	−7.4	−152.8	−208.5	−141.2	−9.6
DO1 (mg/L)	14.6	10.86	8.94	10.34	11.29	7.73	8.01	3.05	4.43	8.95
EC1uS/cm@25°C	63,017	52,893	31,849.67	65,333	35,218	63,792	64,537	53,450	43,879	58,426
RES1 (Ohms·cm)	16	19	22	16	30	16	17	20	25	18
TDS1 (mg/L)	40,330	33,851	42,860.67	41,813	22,539	40,826	41,303	34,208	28,082	37,392
SAL1 (PSU)	42.52	34.84	22.153	44.29	22.16	43.1	43.7	35.26	28.27	39
SSG1 (st)	30.9	24.9	14.7	32.1	15.1	31.3	32.3	25.5	20.2	28.4
TN1 (mg/L)	50.4	64.4	68.6	49	70	44.8	135.8	60.2	162.4	65.8
COD1 (mg/L)	475	2600	1474.333	125	2550	575	123	2070	1150	1425
Total pH.1 (mg/L)	0.079	0.009	0.19	0.053	0.083	0.233	0.033	0.033	0.009	0.056
cd1	0.039	0.037	0.023	0.037	0.041	0.039	0.032	0.03	0.037	0.037
cu1	0.079	0.074	0.062	0.074	0.083	0.085	0.071	0.068	0.073	0.076
zn1	0.062	0.033	0.033	0.032	0.036	0.041	0.031	0.037	0.04	0.038
pb1	0.012	0.007	0.007	0.006	0.007	0.01	0.006	0.007	0.008	0.008
fe1	0.729	0.487	0.529	0.559	1.038	0.986	0.475	0.342	0.901	0.472

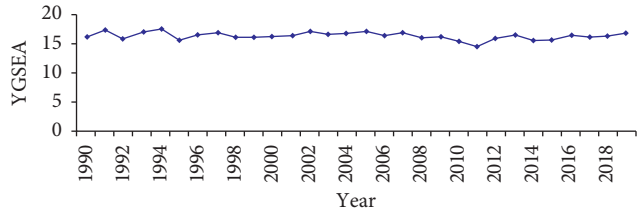

Figure 2: Yearly global solar energy average in Queensland, Australia.

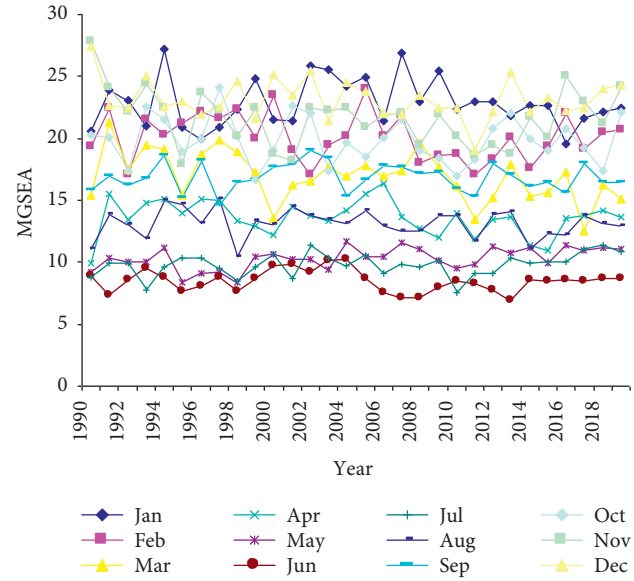

Figure 3: Monthly global solar energy average in Queensland, Australia.

and adjusted R-square. In effect, these procedures produced high accurate estimates of water elements at various depths, and a computer program was created to calculate them.

2.3. Complexity Solution. For this reason, the algorithm of estimation regression models (AERMs) was created to estimate the regression model. It was implemented on the

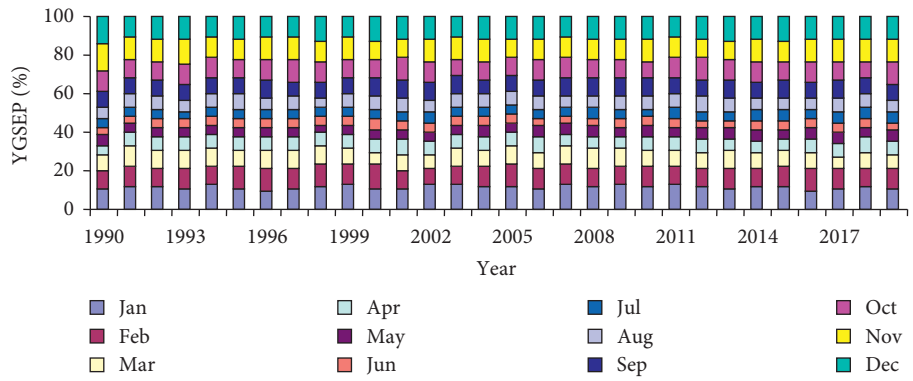

Figure 4: Yearly global solar energy percentage in Queensland, Australia.

computer to evaluate the parameters of regression models. Consequently, the values of water elements in the different depths of the lake were obtained with high accuracy.

2.4. Why Data Set of Solar Energy May Be Complex? Many data sets have high dispersion or big size such as data set of solar energy through years. For this reason, we supposed that $X(t)$ is a stochastic process, where X is a continuous random variable which represents the amount of solar energy with a discrete time, t representing the day. Markov transition probability p_{ij} represents transferring the state i to the state j, whereby $X(t)$ satisfies Markov property.

$$
\begin{aligned}
&P\left(X_{t+1}=j \mid X_t i_t, X_{t-1}=i_{t-1}, \ldots, X_0=i_0\right)=P\left(X_{t+1}=j \mid X_0=i_0\right)=P_{i,j},\\
&P\left(X_{t+1}=j \mid X_t=i_t,\ X_{t-1}=i_{t-1}, \ldots, X_0=i_0\right)\\
&=\frac{p\left(X_{t+1} \cap X_0 \cap X_1 \cap \cdots \cap X_{t-2} \cap X_{t-1} \cap X_t\right)}{p\left(X_0, X_1, \ldots, X_t\right)},\\
&=\frac{P\left(X_{t+1} \mid X_t\right).p\left(X_{t-1}\right).p\left(X_{t-2}\right), \ldots, p\left(X_1\right).p\left(X_0\right)}{p\left(X_0\right).p\left(X_1\right), \ldots, p\left(X_{t-1}\right)},\\
&=P\left(X_{t+1} \mid X_t\right)=P\left(X_{t+1}=j \mid X_t=i_t=i\right)=P_{ij}.
\end{aligned}
\tag{1}
$$

Thus, X_{t+1} depends only on the current state X_t, without knowing all history of states $X_0, X_1, \ldots, X_{t-1}$. Transition matrix P shows the transitions between different states. It was used for forecasting the monthly global solar energy (MGSE).

$$
P=\begin{bmatrix}
X_{(1,1)} & X_{(1,2)} & \cdots\cdots & \cdots\cdots & X_{(1,12)}\\
X_{(2,1)} & X_{(2,2)} & \cdots\cdots & \cdots\cdots & X_{(2,12)}\\
\vdots & \vdots & & \vdots & \vdots\\
\vdots & \vdots & & \vdots & \vdots\\
X_{(12,1)} & X_{(12,2)} & \cdots\cdots & \cdots\cdots & X_{(12,12)}
\end{bmatrix}.
\tag{2}
$$

The probability of the predicted values for the 12 months of next year is by multiplying the vector of current states by the transition matrix P, whereby Y presents the vector of the probabilities of states in the next year.

$$
\begin{aligned}
Y=&\begin{bmatrix} X_1 & X_2 & \ldots & X_{12} \end{bmatrix}\\
&\cdot\begin{bmatrix}
X_{(1,1)} & X_{(1,2)} & \cdots\cdots & \cdots\cdots & X_{(1,12)}\\
X_{(2,1)} & X_{(2,2)} & \cdots\cdots & \cdots\cdots & X_{(2,12)}\\
\vdots & \vdots & & \vdots & \vdots\\
\vdots & \vdots & & \vdots & \vdots\\
X_{(12,1)} & X_{(12,2)} & \cdots\cdots & \cdots\cdots & X_{(12,12)}
\end{bmatrix}.
\end{aligned}
\tag{3}
$$

In this study, a new statistical method called converting data set to Markov model (CDMM) is produced to give an accurate prediction of the monthly average of global solar exposure giving the probabilities transition matrix. Software programs Mathematica, SPSS, and Excel were used to fulfill the research requirements, evaluate the calculations, and extract the numerical results. In addition, nonparametric tests were applied for validity results with the actual data.

2.5. Processing Missing Data Set of Solar Energy. There was a missing value in the month of December, 2005. Through the method of regression for December values for 30 years with the principle of Markov, the future value depends only on the current value; two points before the missing point and two points after it was used to estimate the approximating curve to predict the missing point as shown in Figure 5. The best-fitting curve of the data set in December is shown in Figure 5.

Consequently, the nonlinear regression equation ($y = 0.4583x^3 - 22.536x^2 + 367.43x - 1963.4$) was obtained, whereby R-square = 1, x is the rank of the year in 26 years from 1990 to 2019, and y is the global solar exposure in December. Therefore, the missing estimated value of y at $x = 16$ is equal to 23.8 for 2005.

2.6. Conversion Data Set of Solar Energy into Markov Model (CDMM). Suppose that $\overline{T}_{i,j}$ represents the average temperature for a month (i), $i = 1,2,\ldots, 12$, in the year (j), $j = 1990, 1991,\ldots, 2019$. The states of the temperature $\overline{T}_{i,j}$ for each month (i) are low ($a \le \overline{T}_i < b$), medium ($b \le \overline{T}_i < c$), and high ($\overline{T}_i \ge c$), where a, b, and c are constant temperatures for all months. Then, the number of the transition states of the average temperature $\overline{T}_{i,j}$ for each month (i) can be calculated separately from 1990 to 2019. Thus, the probabilities of transition states and the probability of each state for each month (i) during 1990 to 2019 were obtained, respectively. As a result, the electrical energy corresponding to the cases of solar radiation is known, and thus the number of units benefiting from this electrical energy is estimated.

For this purpose, data sets can be transformed to fit the Markov model by classifying the solar energy to low (L), medium (M), and high (H). Consequently, there are nine transition probabilities: $\{p(X_i), p(Y_i), p(Z_i); i = 1, 2, 3\}$ among three states L, M, and H.

$$\begin{pmatrix} X_1: L \Rightarrow L \longrightarrow n \cdot (LL) \\ X_2: L \Rightarrow M \longrightarrow n \cdot (LM) \\ X_3: L \Rightarrow H \longrightarrow n \cdot (LH) \end{pmatrix} \mapsto \begin{pmatrix} p(X_1) = \dfrac{n \cdot (LL)}{n \cdot (LL) + n \cdot (LH) + n \cdot (LM)} \\ p(X_2) = \dfrac{n \cdot (LM)}{n \cdot (LL) + n \cdot (LH) + n \cdot (LM)} \\ p(X_3) = \dfrac{n \cdot (LH)}{n \cdot (LL) + n \cdot (LH) + n \cdot (LM)} \end{pmatrix},$$

$$\begin{pmatrix} Y_1: M \Rightarrow M \longrightarrow n \cdot (MM) \\ Y_2: M \Rightarrow L \longrightarrow n \cdot (ML) \\ Y_3: M \Rightarrow H \longrightarrow n \cdot (MH) \end{pmatrix} \mapsto \begin{pmatrix} p(Y_1) = \dfrac{n \cdot (MM)}{n \cdot (LL) + n \cdot (LH) + n \cdot (LM)} \\ p(Y_2) = \dfrac{n \cdot (ML)}{n \cdot (LL) + n \cdot (LH) + n \cdot (LM)} \\ p(Y_3) = \dfrac{n \cdot (MH)}{n \cdot (LL) + n \cdot (LH) + n \cdot (LM)} \end{pmatrix}, \tag{4}$$

$$\begin{pmatrix} Z_1: H \Rightarrow H \longrightarrow n \cdot (HH) \\ Z_2: H \Rightarrow L \longrightarrow n \cdot (HL) \\ Z_3: H \Rightarrow M \longrightarrow n \cdot (HM) \end{pmatrix} \mapsto \begin{pmatrix} p(Z_1) = \dfrac{n \cdot (HH)}{n \cdot (HH) + n \cdot (HL) + n \cdot (HM)} \\ p(Z_2) = \dfrac{n \cdot (HL)}{n \cdot (HH) + n \cdot (HL) + n \cdot (HM)} \\ p(Z_3) = \dfrac{n.(HM)}{n \cdot (HH) + n \cdot (HL) + n \cdot (HM)} \end{pmatrix}.$$

Then,

$$P = \begin{matrix} \\ L \\ M \\ H \end{matrix} \begin{matrix} \begin{matrix} L & M & H \end{matrix} \\ \begin{bmatrix} p(X_1) & p(Y_2) & p(Z_2) \\ p(X_2) & p(Y_1) & p(Z_3) \\ p(X_3) & p(Y_3) & p(Z_1) \end{bmatrix} \end{matrix}. \tag{5}$$

Figure 6 shows that Markov model for the transition states of L, M, and H, whereby the summation of the probabilities of each column in the matrix P is equal to 1. Figure 7 demonstrates the Markov model for six states L, $L+$, M, $M+$, $H+$, and 36 transition states, where $L+$, $M+$, and $H+$ represent the states of solar energy above low, above medium, above high, respectively.

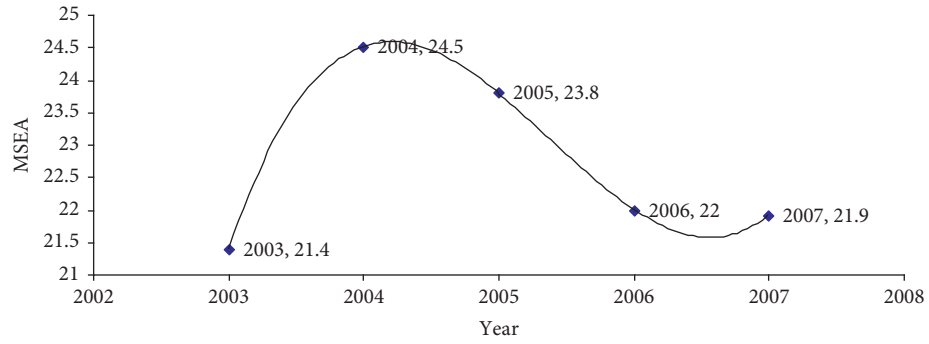

Figure 5: Best-fitting regression curve of MSEA in December, Queensland, Australia.

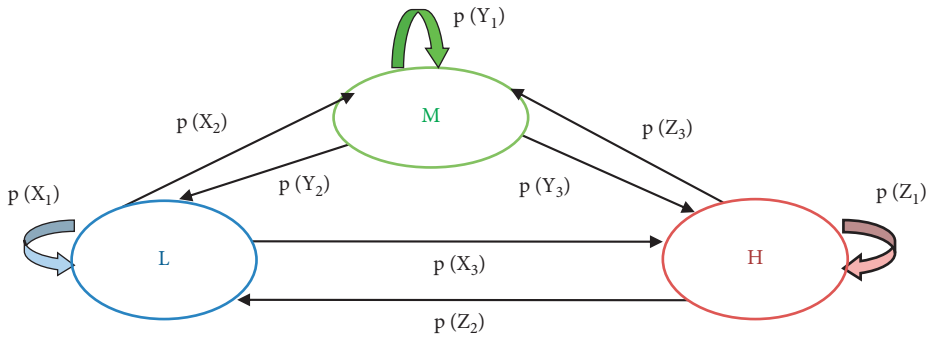

Figure 6: Markov model of the solar energy with three states.

2.7. Joint Probability Function Combined with the Property and Process of Markov Model. Suppose that there are four different transition states for a month in two consecutive years: V_I, V_J, V_K, and V_L and this is for two phenomena: solar radiation (S) and corresponding temperature (T) simultaneously. Both X and Y are continuous random variables represent S and T, respectively, which satisfy the conditions of Markov process and Markov property. The increasing, decreasing, and unchanged amount of the solar radiation for a month in two consecutive years can be represented by X^+, X^-, and X^o, respectively. Similarly, Y^+, Y^-, and Y^o represent the increasing, decreasing, and unchangeable temperatures for a month in two consecutive years, respectively.

Let $\overline{D}_S$ refer to the average of different values of X, and $\overline{D}_T$ stand for the average of the different values of Y. The four different transition states for a month in two consecutive years can be defined as follows:

State I:

$$p(V_I) = p\left(X < x \pm \overline{D}_S | Y < y \pm \overline{D}_T\right) \cdot p\left(Y < y \pm \overline{D}_T\right) \cdot \underset{\overline{D}_S}{p}\left(x^{\mp}\right) \cdot \underset{\overline{D}_T}{p}\left(y^{\mp}\right). \tag{6}$$

State J:

$$p\left(V_J\right) = p\left(X < x | Y < y \pm \overline{D}_T\right) \cdot p\left(Y < y \pm \overline{D}_T\right) \cdot \underset{\overline{D}_S}{p}\left(x^{o}\right) \cdot \underset{\overline{D}_T}{p}\left(y^{\mp}\right). \tag{7}$$

State K:

$$p\left(V_K\right) = p\left(X < x \pm \overline{D}_S | Y < y\right) \cdot p\left(Y < y\right) \cdot \underset{\overline{D}_S}{p}\left(x^{\mp}\right) \cdot \underset{\overline{D}_T}{p}\left(y^{o}\right). \tag{8}$$

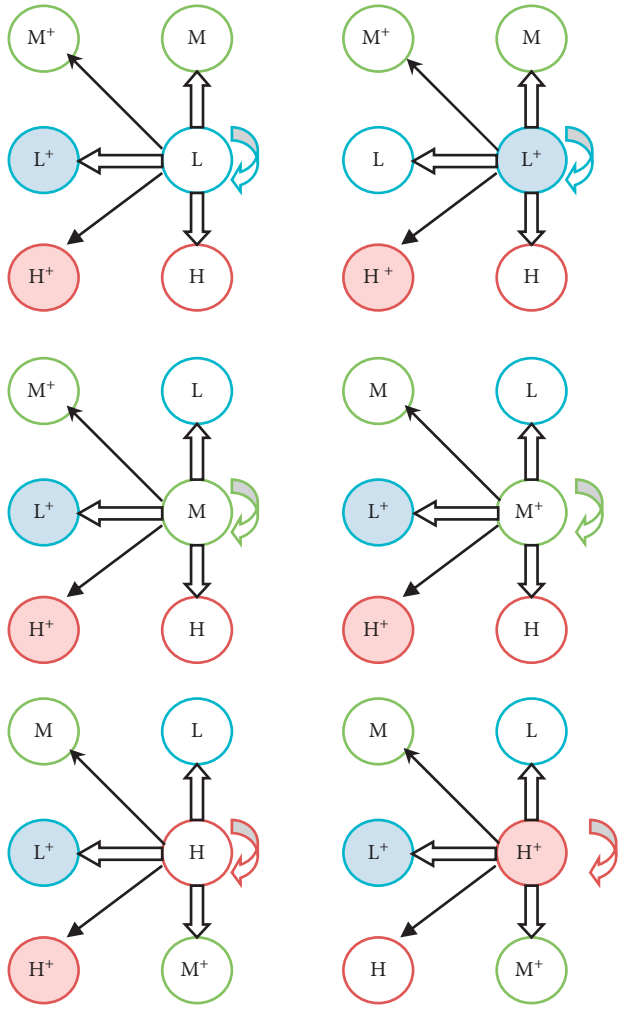

Figure 7: Markov model of the solar energy for the states (*L*) *L*+, (*M*) *M*+, and *H*+.

State *L*:

$$p(V_L) = p(X < x \mid Y < y) \cdot p(Y < y) \cdot \underset{\overline{D}_S}{p}(x^o) \cdot \underset{\overline{D}_T}{p}(y^o). \quad (9)$$

Then, the joint probability function $F(x, y)$ will have the following equation:

$$\begin{aligned} F(X < x, Y < y) &= p(V_I) + p(V_J) + p(V_K) + p(V_L) \\ &= p(X < x \pm \overline{D}_S \mid Y < y \pm \overline{D}_T) \cdot p(Y < y \pm \overline{D}_T) \cdot \underset{\overline{D}_S}{p}(x^{\mp}) \cdot \underset{\overline{D}_T}{p}(y^{\mp}), \\ & p(X < x \mid Y < y \pm \overline{D}_T) \cdot p(Y < y \pm \overline{D}_T) \cdot \underset{\overline{D}_S}{p}(x^o) \cdot \underset{\overline{D}_T}{p}(y^{\mp}), \\ & p(X < x \pm \overline{D}_S \mid Y < y) \cdot p(Y < y) \cdot \underset{\overline{D}_S}{p}(x^{\mp}) \cdot \underset{\overline{D}_T}{p}(y^o), \\ & p(X < x \mid Y < y) \cdot p(Y < y) \cdot \underset{\overline{D}_S}{p}(x^o) \cdot \underset{\overline{D}_T}{p}(y^o), \end{aligned} \quad (10)$$

whereas I, J, K, and L signify all the different transition states of X and Y for a month in two consecutive years, as follows:

$$\begin{aligned} I\colon (X^{\pm}, Y^{\pm}) &\longrightarrow (X^{\mp}, Y^{\mp}), (X^{\pm}, Y^{\mp}) \longrightarrow (X^{\mp}, Y^{\pm}),\\ J\colon \left(X^{0}, Y^{\pm}\right) &\longrightarrow \left(X^{0}, Y^{\mp}\right),\\ K\colon \left(X^{\pm}, Y^{0}\right) &\longrightarrow \left(X^{\mp}, Y^{0}\right),\\ L\colon \left(X^{0}, Y^{0}\right) &\longrightarrow \left(X^{0}, Y^{0}\right). \end{aligned} \tag{11}$$

Moreover, the probabilities $\underset{\overline{D}_S}{p}(x^{-})$, $\underset{\overline{D}_T}{p}(y^{-})$, $\underset{\overline{D}_S}{p}(x^{+})$, $\underset{\overline{D}_T}{p}(y^{+})$, $\underset{\overline{D}_S}{p}(x^{o})$, and $\underset{\overline{D}_T}{p}(y^{o})$ are defined as shown below. $\underset{\overline{D}_S}{p}(x^{-})$ is the probability when x is decreased by $\overline{D}_S$, $\underset{\overline{D}_S}{p}(y^{-})$ is the probability when y is decreased by $\overline{D}_S$, $\underset{\overline{D}_S}{p}(x^{+})$ is the probability when x is increased by $\overline{D}_S$, $\underset{\overline{D}_T}{p}(y^{+})$ is the probability when y is increased by $\overline{D}_T$, $\underset{\overline{D}_S}{p}(x^{o})$ is the probability when x is unchangeable, and $\underset{\overline{D}_T}{p}(y^{o})$ is the probability when y is unchanged.

Remarkably, if X and Y had the exponentiated Gumbel maximum distribution, then the probability density functions and cumulative distribution functions would be as follows:

$$\begin{aligned} f(Z=z) &= \frac{\mu_\gamma}{\beta_\gamma} e^{-z/\beta_\gamma} e^{-\mu_\gamma e^{-z/\beta_\gamma}}, \quad z \in (-\infty; +\infty),\\ F(Z=z) &= e^{-\mu_\gamma e^{-z/\beta_\gamma}}, \quad z \in (-\infty; +\infty), \end{aligned} \tag{12}$$

where $Z = \{X, Y\}$ and corresponding $\gamma = \{S, T\}$, respectively. It is worth mentioning that $\beta_S > 0$ and $\beta_T > 0$ are scale parameters and $\mu_S > 0$ and $\mu_T > 0$ are shape parameters.

3. Results and Discussion

3.1. Methodology and Results of AERM Method. It focuses on the points of contribution that hopefully foreground the significance of this study in terms of homogeneity test and Mann–Whitney test to compare the values of the actual data set with the estimation values by AERM method. Tables 2 and 3 show that the values of the estimated parameters in the regression equations at various depths of less-estimate error (Algorithm 1).

The AERM algorithm aims to read the points of the scales data for any sample. The sample contains several items with different measurements, even if there is high dispersion between the measurement values for each item separately. The AERM algorithm also handles large or small sample sizes that contain several elements. For this reason, the AERM algorithm overcame the problem of the difficulty of the accuracy of predicting the measurements of each element in the different intervals of the depths of Lake Manzala in Egypt.

Lake Manzala can be divided into two main regions according to their salinities. First, the southern region of the lake that is characterized by lower values of salinity and high concentration of nutrients and heavy metal as a consequence of receiving high volumes of low salinity drainage water through various drains. The second is the region at the north eastern area of the lake, near the lake-sea connection (El-Gamil), which is characterized by high salinity values and low nutrient concentration as a result of seawater intrusion through the outlet openings. Sallam and Elsayed discovered that the lake is exposed to high levels of pollutants from industrial, domestic, and agricultural sources [21].

It has increasingly been subject to human pressures, including rapid municipal growth at Port Said, Damietta, and El Mataria, and reduction in lake surface by illegal land reclamation of wetlands for agriculture. Barakat et al. asserted that it produces about 50% of the fish catch of the northern lakes and freshwater fisheries [22].

The results of the implementation of the AERM algorithm are charts of the linear and nonlinear regression of the water elements in Lake Manzala, which are shown in Figures 8 to 12. Also, the parameter estimates for the linear and nonlinear regression lines and their equations are shown in Tables 2 and 3.

In effect, it is difficult to predict the maximum number of migratory bird types during a limited number of migration years, especially because of the large number of bird types in the world, a multiple nonlinear regression model of maximum number probability function of migratory bird types was obtained by El Genidy [23]. On the contrary, a multiple nonlinear regression model could accurately perform the job prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques, as such presented by Gong and Ordieres-Mere [24].

In this study, however, the homogeneity test and Mann–Whitney test were applied to obtain the real values along with the estimated values, with the use of the significant proposed program. Consequently, the two groups of values are compared to evaluate both the applicability and the accuracy of the program. In addition, the arithmetic mean and difference factor were also calculated to compare the averages of element sample values and their corresponding dispersion measures. The statistical significance p values precisely reveal that no essential differences are foregrounded. This means that the newly proposed program efficiently works.

The linear and nonlinear regressions for 17 elements in different depth intervals of Lake Manzala in Egypt shown in Figures 8 to 12 clarify the accuracy of AERM's algorithm in predicting with least error $\mathrm{SSE} \leq 0.02$ and $R^2 \geq 0.98$.

Thus, it is possible to predict the unknown values of any elements in water using polynomial equations arising from the AERM's algorithm in Tables 2 and 3. It should be noted that AERM is a general algorithm for any other samples in different applications because their inputs of the values of variables can be changed according to application data.

Table 4 demonstrates the estimated values of the water elements at various depths ranging from 0.13 to 2.5 m by the estimator program in Lake Manzala, Egypt. Table 5 shows the comparison between the estimated values of the elements of water at various depths in using coefficient of variation,

Table 2: Regression model polynomial of the second degree.

Model	Depth	Regression equation
$y = ax^2 + bx + c$	$0.13 \le x < 0.31$	Temp (2) = $368.68x^2 - 154.22x + 35.918$
		pH (2) = $-797.13x^2 + 334.61x - 21.488$
		ORP (2) = $68081x^2 - 27968x + 2434.7$
		DO (2) = $-991.91x^2 + 393.23x - 19.756$
		RES (2) = $1654.4x^2 - 676.1x + 75.934$
		SAL (2) = $-7180x^2 + 2967.8x - 221.95$
		SSG (2) = $-5779.4x^2 + 2389.9x - 182.11$
		TN (2) = $1955.9x^2 - 733.97x + 112.76$
		cd (2) = $-8.1618x^2 + 3.4154x - 0.2671$
		cu (2) = $-6.875x^2 + 2.8562x - 0.1761$
		zn (2) = $1.0662x^2 - 0.629x + 0.1258$
		pb (2) = $0.1838x^2 - 0.1085x + 0.023$
		fe (2) = $33.603x^2 - 15.626x + 2.1925$
	$0.31 \le x < 1.34$	Temp (2) = $714.29x^2 - 497.14x + 108.07$
		pH (2) = $-491.43x^2 + 334.51x - 47.783$
		ORP (2) = $18814x^2 - 13666x + 2469$
		DO (2) = $-1695x^2 + 1132.8x - 177.86$
		RES (2) = $-14000x^2 + 9660x - 1633.2$
		SAL (2) = $21790x^2 - 15052x + 2616.4$
		SSG (2) = $16771x^2 - 11584x + 2011.3$
		TN (2) = $-22200x^2 + 15258x - 2547.6$
		cd (2) = $-3.4286x^2 + 2.3943x - 0.3757$
		Cu (2) = $-5.8571x^2 + 4.1986x - 0.6647$
		zn (2) = $-1.4286x^2 + 1.1143x - 0.1761$
		pb (2) = $0.1429x^2 - 0.0414x + 0.0051$
		fe (2) = $-357x^2 + 252.43x - 43.387$
	$0.3 \le x < 0.38$	Total pb (2) = $506.67x^2 - 322.77x + 51.42$

(2) refers to the estimated elements in Lake Manzala by regression model.

Table 3: Regression models polynomial of the first, second, and third degree.

Model	Depth	Regression equation
$y = ax^3 + bx^2 + cx + d$	$1.34 \le x \le 2.5$	Temp (2) = $9.5611x^3 - 55.367x^2 + 104.34x - 43.508$
		pH (2) = $-3.854x^3 + 23.964x^2 - 47.844x + 38.775$
		ORP (2) = $-519.66x^3 + 3246.5x^2 - 6421.5x + 3872.8$
		DO (2) = $-38.682x^3 + 238.37x^2 - 473.74x + 307.87$
		RES (2) = $-9.2607x^3 + 31.564x^2 - 14.813x + 2.4553$
		SAL (2) = $-20.088x^3 + 155.97x^2 - 374.06x + 313.21$
		SSG (2) = $-18.391x^3 + 137.34x^2 - 321.19x + 260.33$
		TN (2) = $-1226.9x^3 + 6963.8x^2 - 12820x + 7762.6$
		cd (2) = $-0.0467x^3 + 0.2661x^2 - 0.4855x + 0.3171$
		cu (2) = $-0.041x^3 + 0.242x^2 - 0.4578x + 0.3485$
		zn (2) = $0.0325x^3 - 0.2051x^2 + 0.4232x - 0.2459$
		pb (2) = $0.0038x^3 - 0.025x^2 + 0.054x - 0.0307$
		fe (2) = $-4.2495x^3 + 23.521x^2 - 41.898x + 24.608$
	$0.38 < x \le 2.5$	Total pb (2) = $0.006x^3 + 0.076x^2 - 0.3468x + 0.3533$
$y = ax + b$&$y = ax^2 + bx + c$	$0.13 \le x < 0.3$	Total pb (2) = $-0.4375x + 0.1359$
	$0.13 \le x < 0.3$	EC (2) = $-63275x + 71243$
	$0.3 \le x < 0.38$	EC (2) = $-2E + 08x^2 + 1E + 08x - 2E + 07$
	$0.38 \le x < 2$	EC (2) = $-62562x^2 + 10838x + 31641$
	$2 \le x \le 2.5$	EC (2) = $29094x - 14309$
	$0.13 \le x < 0.3$	COD (2) = $13281x - 1251.6$
	$0.3 \le x < 0.38$	COD (2) = $9E + 06x^2 - 5E + 06x + 83615$
	$0.38 \le x < 2$	COD (2) = $12169x - 16183$
	$2 \le x \le 2.5$	COD (2) = $550x + 50$
	$0.13 \le x < 0.3$	TDS (2) = $-40494x + 45594$
	$0.3 \le x < 0.38$	TDS (2) = $-3E + 07x^2 + 2E + 07x - 3E + 06$
	$0.38 \le x < 2$	TDS (2) = $-40036x^2 + 69359x + 20251$
	$2 \le x \le 2.5$	TDS (2) = $18620x - 9158$

(1) Input n "n is the number of elements in the sample."
(2) Input M "where $M > 4$ is the total number of data points for the element (i)."
(3) For $i = 1$ to n.
(4) let $Q = M : r = 1$.
(5) For $j = r$ to M.
(6) For $k = 4$ to 2 Step 1 "k is the number of data points."
(7) If $k \le Q$ then read k points (X_j, Y_j) from data.
(8) Estimate the model of nonlinear regression $Y = \sum_{L=0}^{k-1} a_L X^L$ form the determinant:

$$\begin{vmatrix} Y & X^{k-1} & X^{k-2} & X & 1 \\ \sum_{j=r}^{r+k-1} Y_j & \sum_{j=r}^{r+k-1} X_j^{k-1} & \sum_{j=r}^{r+k-1} X_j^{k-2} & \sum_{j=r}^{r+k-1} X_j & k \\ \sum_{j=r}^{r+k-1} X_j Y_j & \sum_{j=r}^{r+k-1} X_j^{k-2} & \sum_{j=r}^{r+k-1} X_j^{k-3} & \sum_{j=r}^{r+k-1} X_j^2 & \sum_{j=r}^{r+k-1} X_j \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{j=r}^{r+k-1} X_j^{k-2} Y_j & \sum_{j=r}^{r+k-1} X_j^{2k-3} & \sum_{j=r}^{r+k-1} X_j^{2k-4} & \sum_{j=r}^{r+k-1} X_j^{k-1} & \sum_{j=r}^{r+k-1} X_j^{k-2} \\ \sum_{j=r}^{r+k-1} X_j^{k-1} Y_j & \sum_{j=r}^{r+k-1} X_j^{2k-2} & \sum_{j=r}^{r+k-1} X_j^{2k-3} & \sum_{j=r}^{r+k-1} X_j^{k} & \sum_{j=r}^{r+k-1} X_j^{k-1} Y_j \end{vmatrix} = 0$$

(9) Compute R-square (R_2), the total sum of squares (SST), the regression sum of squares (SSR), and the error sum of squares (SSE) from: $\mathrm{SSR} = \sum_{j=r}^{r+k-1} (\hat{Y}_j - \overline{Y_j})^2$, $\mathrm{SSE} = \sum_{j=r}^{r+k-1} (\hat{Y}_j - Y_j)^2$, $\mathrm{SST} = \sum_{j=r}^{r+k-1} (Y_j - \overline{Y})^2$, where $\overline{Y} = \sum_{j=r}^{r+k-1} Y_j / k$, $R^2 = \mathrm{SSR}/\mathrm{SST}$. $\hat{Y}_j$ is the estimated value from the equation $Y = \sum_{L=0}^{k-1} a_L X^L$ by the determinant in the Step 8, and Y_j is the actual data of the element (i).
(10) If $R^2 \ge 0.95$ then: {print the equation in Step 8, $r = r + k$, $Q = Q - r + 1$, Return to the loop in Step 5}, else: Return to the loop in Step 6 where k will be equal to $k - 1$.
(11) When $k = 2$, estimate the simple linear regression equation $Y = \sum_{L=0}^{k-1} a_L X^L$ similarly as in the determinant of Step 8. As well $r = r + k$, $Q = Q - r + 1$ then return to Step 5.
(12) If $r = M$ or $Q = Q - M + 1$ then: {Return to the loop in the Step 3 and move to the next element $(i + 1)$}.
(13) The algorithm ends after the loop in Step 3 has finished; thus, all prediction equations are obtained for all elements in the given sample.

Algorithm 1: Algorithm of estimation regression models (AERMs).

averages, standard deviation, homogeneity test, and Mann–Whitney test.

3.2. Validity of AERM Method. As a result, the high accurate parameters estimation method by AERM was performed to estimate the elements of water quality at various depths in Lake Manzala, Egypt. In addition, AERM technique achieved all required results.

Referring to related estimation methods, multiple linear regression, and artificial neural networks, based on principal components, could predict ozone concentrations. Later on, an adaptive neuro fuzzy inference system to anticipate the ground inflow into Amir Kabir tunnel in Iran, whereby a sample of 110 data sets containing the most influential parameters on ground inflow rate was set to develop the ground inflow rate forecasting model, introduced by Mecibah et al. [25]. Typical of the results of some previous studies is that the exponentiated Gumbel maximum distribution was estimated by quartiles moments method to analyze the maximum temperature and solar radiation data, thence, multiple nonlinear regressions of the daily global solar radiation, and the corresponding daily maximum temperatures are produced and compared with the real data set accordingly, carried out by El Genidy [26].

It is necessary to obtain a large amount of sensor data correctly and effectively. A new method has been proposed to filter noise from the input sensor called Kalman filtering, which is one of the most representative filtering techniques. Kalman filtering corrects inaccurate values of input sensor data [27].

3.3. Applying CMMD Method to Solar Energy. In the case of three states L, M, and H, the data sets of solar energy in January 2019 were divided into three half-open intervals [1–12), [12–23), and [23–34). These intervals are corresponding to the states L, M, and H, respectively. After converting the data set into Markov model, the transition probability matrix T of the three states and the vector V_c of the probabilities of the current state in January 2019 were obtained.

$$T = \begin{array}{c} \\ L \\ M \\ H \end{array} \begin{array}{c} \begin{array}{ccc} L & M & H \end{array} \\ \begin{bmatrix} \frac{2}{6} & \frac{2}{8} & \frac{2}{15} \\ \frac{2}{6} & 0 & \frac{5}{15} \\ \frac{2}{6} & \frac{6}{8} & \frac{8}{15} \end{bmatrix} \end{array}, \quad V_c = \begin{array}{c} \\ L \\ M \\ H \end{array} \begin{bmatrix} \frac{6}{31} \\ \frac{8}{31} \\ \frac{17}{31} \end{bmatrix}. \tag{13}$$

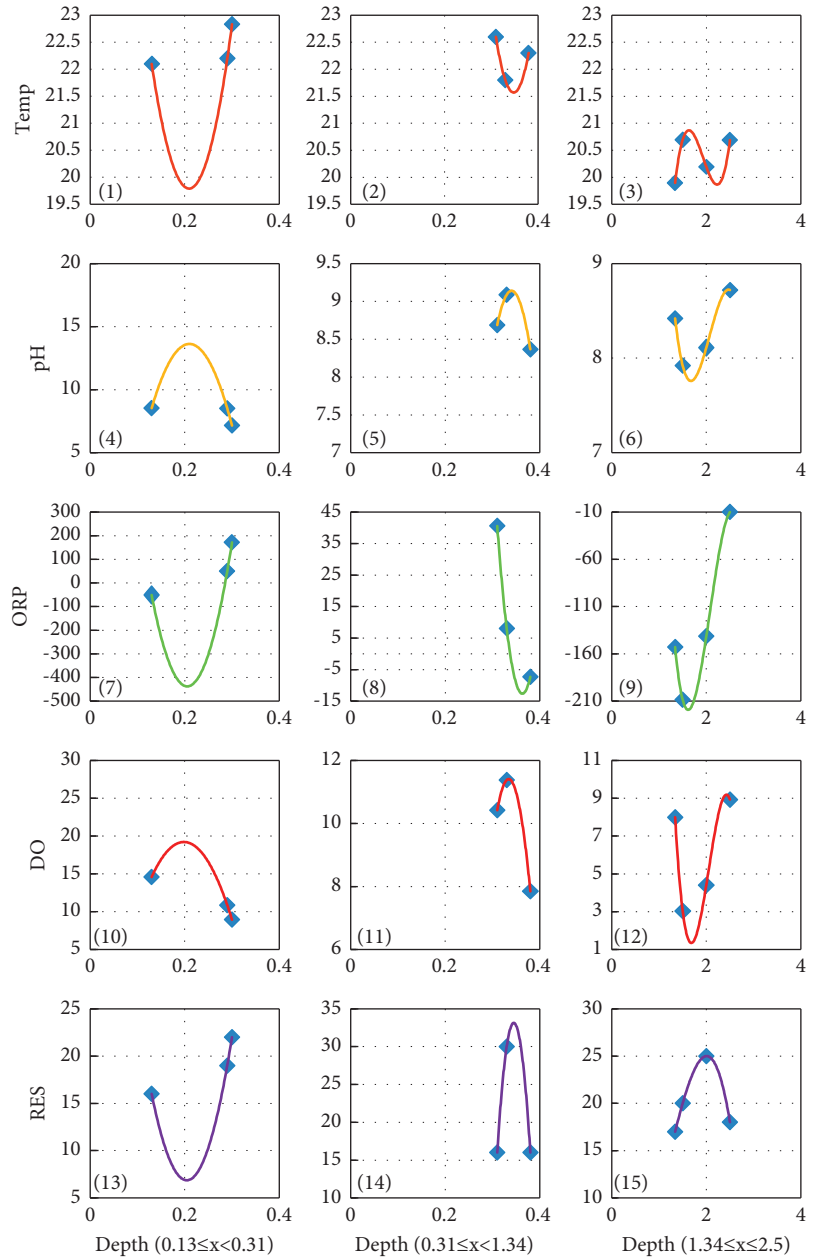

Figure 8: Estimation of the regressions of Temp, pH, ORP, DO, and RES in Lake Manzala, Egypt.

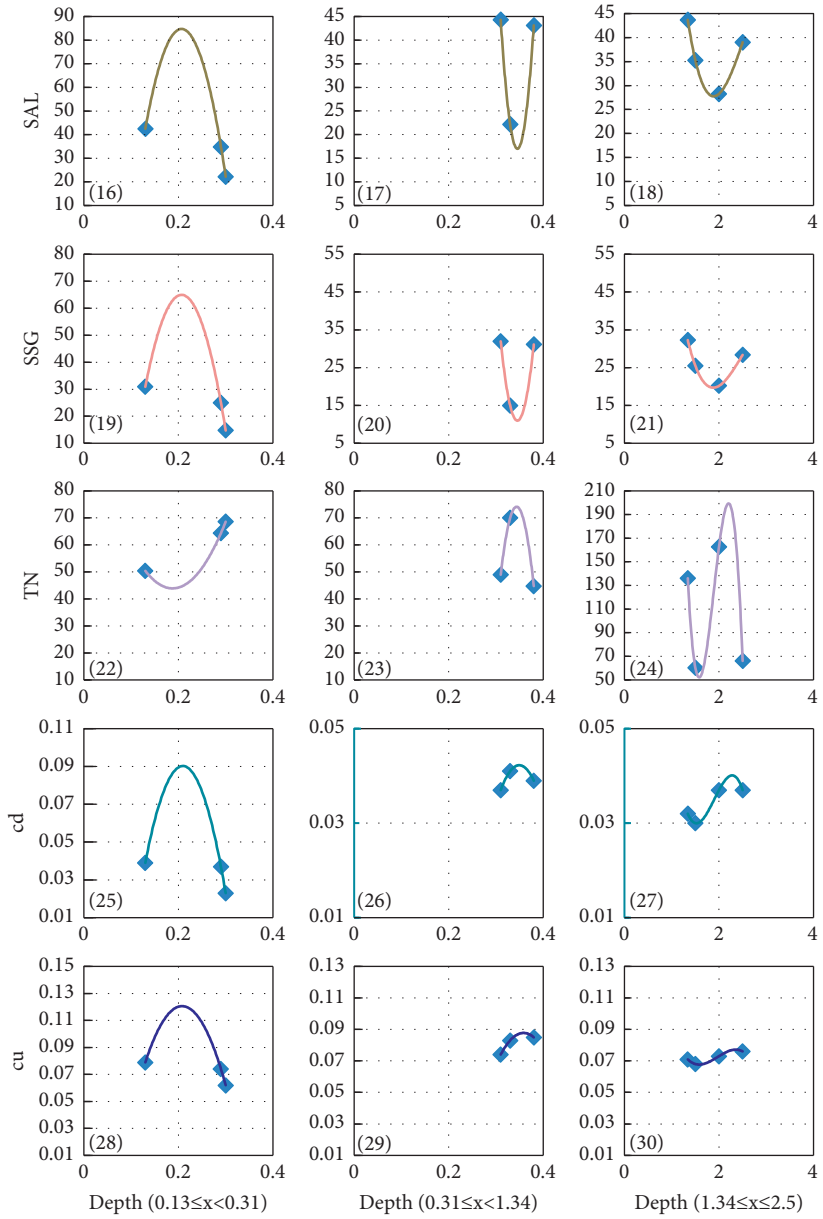

Figure 9: Estimation of the regressions of SAL, SSG, TN, cd, and cu in Lake Manzala, Egypt.

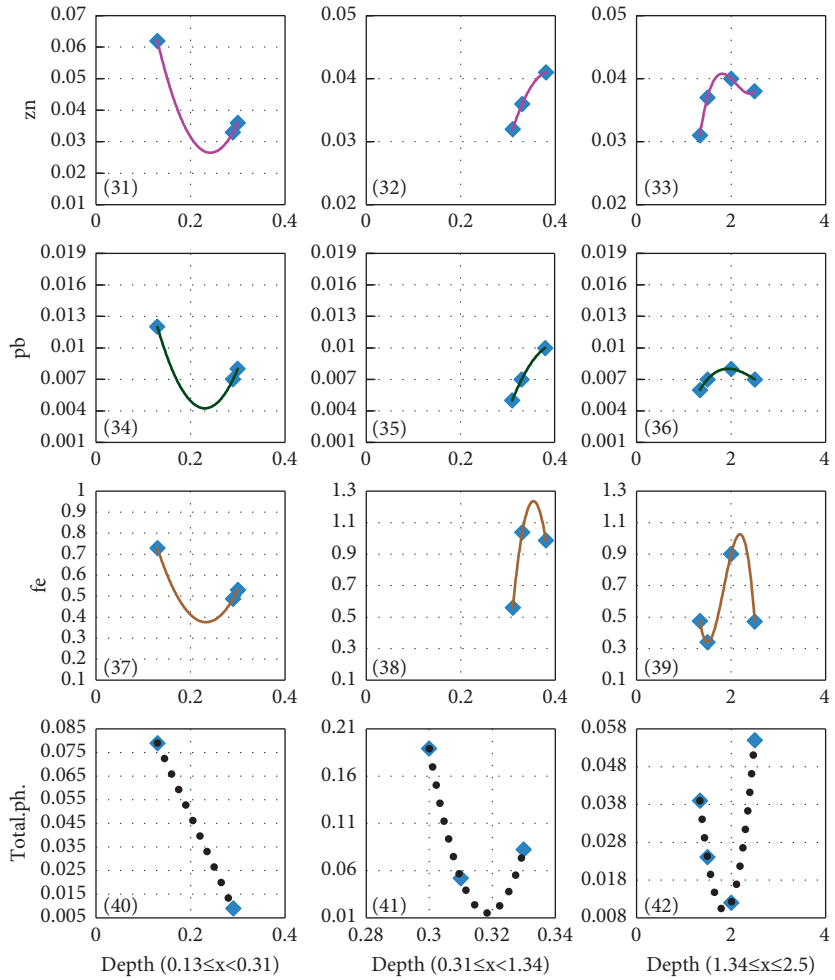

Figure 10: Estimation of the regressions of zn, pb, fe, and total pH in Lake Manzala, Egypt.

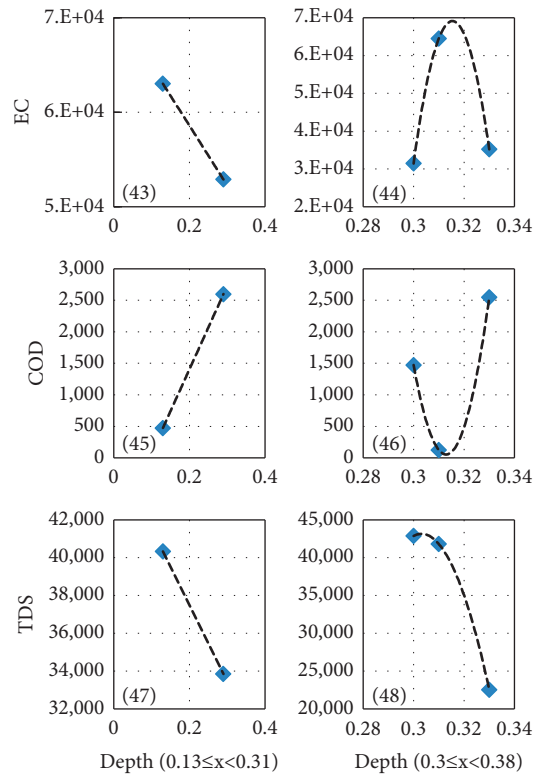

Figure 11: Estimation of the regressions of EC, COD, and TDS for the first two intervals in depth.

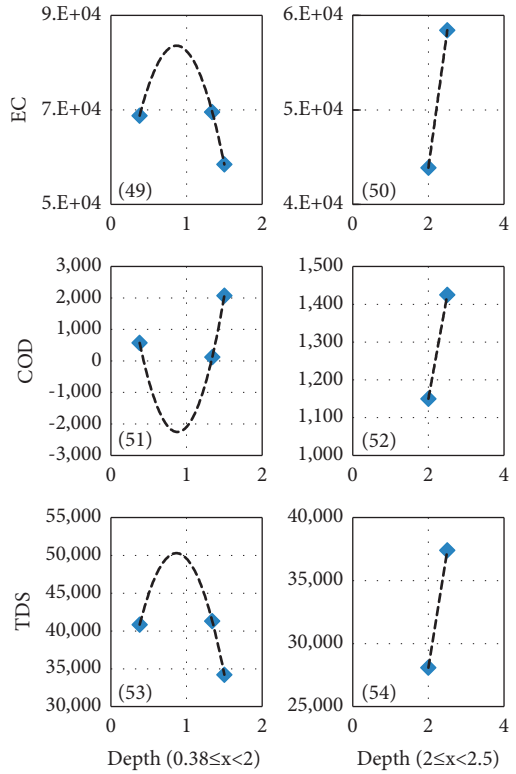

Figure 12: Estimation of the regressions of EC, COD, and TDS for the second two intervals in depth.

Table 4: Estimated values of water elements at different depths of the lake.

Depth (m)	0.13	0.29	0.3	0.31	0.33	0.38	1.34	1.5	2	2.5
Temp2 (°C)	22.1	22.2	22.833	22.6	21.8	22.3	19.896	20.695	20.193	20.69
pH2	8.54	8.51	7.153	8.689	9.089	8.368	8.421	7.921	8.111	8.721
ORP2 (REDOX)	−50.571	49.592	171.59	40.565	8.065	−7.338	−152.951	−208.678	−141.48	−10.013
DO2 (mg/L)	14.6	10.861	8.941	10.419	11.379	7.846	8.003	3.041	4.414	8.926
EC2uS/cm@25°C	63017.25	52893.25	31437.38	64486.47	35218.4	63792.4	64537.04	53450	43879	58426
RES2(Ohms.cm)	16	19	22	16	30	16	17	20	25	18
TDS2 (mg/L)	40329.78	33850.74	42860.9	41813.23	22539.2	40826.2	41303.04	34208	28082	37392
SAL2 (PSU)	42.522	34.874	22.19	44.299	22.171	43.116	43.696	35.256	28.266	38.998
SSG2 (st)	30.905	24.913	14.714	31.953	14.942	31.112	32.292	25.49	20.182	28.371
TN2 (mg/L)	50.399	64.4	68.6	48.96	70	44.76	135.95	60.363	162.6	66.038
COD2 (mg/L)	474.93	2599.89	1474.1	124.77	2550	575	123.46	2070.5	1150	1425
Total pH.2 (mg/L)	0.079	0.009	0.189	0.052	0.082	0.233	0.039	0.024	0.012	0.055
cd2	0.039	0.037	0.023	0.037	0.041	0.039	0.032	0.03	0.037	0.037
cu2	0.079	0.074	0.062	0.074	0.083	0.085	0.071	0.068	0.073	0.076
zn2	0.062	0.033	0.033	0.032	0.036	0.041	0.031	0.037	0.04	0.038
pb2	0.012	0.007	0.007	0.006	0.007	0.01	0.006	0.007	0.008	0.007
fe2	0.729	0.487	0.529	0.559	1.038	0.986	0.474	0.341	0.9	0.471

Table 5: Comparison between the values of estimated water elements and their actual data set.

Water properties	Coefficient of variation (%)	Average	Standard deviation	Homogeneity p value	Mann–Whitney test p value
Temp (1)	4.92	21.5333	1.058605	0.987*	0.820*
Temp (2)	4.93	21.5307	1.061584		
pH (1)	6.36	8.3523	0.531492	0.998*	0.970*
pH (2)	6.36	8.3523	0.531211		
ORP (1)	−375.2	−30.01	112.5976	0.999*	0.880*
ORP (2)	−373.98	−30.1219	112.6491		
DO (1)	37.87	8.82	3.340356	0.999*	0.970*
DO (2)	37.91	8.843	3.352044		
EC (1)	23.24	53239.47	12373.02	0.995*	0.970*
EC (2)	23.28	53113.72	12363.54		

Table 5: Continued.

Water properties	Coefficient of variation (%)	Average	Standard deviation	Homogeneity p value	Mann–Whitney test p value
RES (1)	23.14	19.9	4.605552	1*	1*
RES (2)	23.14	19.9	4.605552		
TDS (1)	18.44	36320.47	6697.048	1*	0.910*
TDS (2)	18.44	36320.51	6697.063		
SAL (1)	24.33	35.5293	8.645166	0.998*	0.880*
SAL (2)	24.31	35.5388	8.639253		
SSG (1)	26.56	25.54	6.784328	0.990*	0.880*
SSG (2)	26.58	25.4874	6.775422		
TN (1)	51.06	77.14	39.39092	0.997*	0.970*
TN (2)	51.1	77.207	39.45543		
COD (1)	74.4	1256.733	934.9629	1*	0.940*
COD (2)	74.39	1256.765	934.9632		
PO4 (1)	97.02	0.0778	0.07548	0.997*	0.940*
PO4 (2)	97.37	0.0774	0.075365		
cd (1)	15.32	0.0352	0.005391	1*	1*
cd (2)	15.32	0.0352	0.005391		
cu (1)	9.15	0.0745	0.006819	1*	1*
cu (2)	9.15	0.0745	0.006819		
zn (1)	23.48	0.0383	0.008994	1*	1*
zn (2)	23.48	0.0383	0.008994		
pb (1)	24.02	0.0078	0.001874	0.971*	0.811*
pb (2)	24.53	0.0077	0.001889		
fe (1)	37.56	0.6518	0.244793	0.998*	0.880*
fe (2)	37.61	0.6514	0.244983		

*indicates the (p value) for each of homogeneity test and Mann–Whitney test.

Then, the vector V_n of the probabilities of the next states in January 2020 is derived from (T. V_c), while V_a is the vector of the probabilities of the actual data set of states in January 2020.

$$V_n = \begin{matrix} L \\ M \\ N \end{matrix}\begin{bmatrix} \frac{94}{465} \\ \frac{23}{93} \\ \frac{256}{465} \end{bmatrix} \cong \begin{matrix} L \\ M \\ N \end{matrix}\begin{bmatrix} 0.20 \\ 0.25 \\ 0.55 \end{bmatrix},$$
$$V_a = \begin{matrix} L \\ M \\ N \end{matrix}\begin{bmatrix} \frac{6}{31} \\ \frac{11}{31} \\ \frac{14}{31} \end{bmatrix} \cong \begin{matrix} L \\ M \\ N \end{matrix}\begin{bmatrix} 0.19 \\ 0.35 \\ 0.46 \end{bmatrix}. \tag{14}$$

It turns out that the absolute error value in predicting the probability of L is equal to 0.01, which is small and acceptable. However, the error values of M and H are approximately equal to 0.1, that is, relatively high error. Reducing error and improving results are done by re-partitioning the data set into six half-open intervals [1,7), [7,13), [13,19), [19,25), [25,31), and [31,34), whereby three new states, above low (L+), above medium (M+), and above high (H+) were added to the previous states Low (L), Medium (M), and High (H). The transition probability matrix S of the states L, L+, M, M+, H, and H+, respectively, and the vector Q_c of the probabilities of the current state in January 2019 were obtained.

$$S = \begin{matrix} \\ L \\ L^+ \\ M \\ M^+ \\ H \\ H^+ \end{matrix}\begin{matrix} \begin{matrix} L & L^+ & M & M^+ & H & H^+ \end{matrix} \\ \begin{bmatrix} 0 & 0 & \frac{1}{4} & 0 & \frac{1}{13} & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{13} & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & 0 & \frac{2}{13} & 0 \\ 0 & 0 & 0 & 1 & \frac{2}{13} & 1 \\ 0 & \frac{2}{4} & \frac{3}{4} & \frac{3}{4} & \frac{7}{13} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix},$$
$$Q_c = \begin{matrix} L \\ L^+ \\ M \\ M^+ \\ H \\ H^+ \end{matrix}\begin{bmatrix} \frac{2}{31} \\ \frac{4}{31} \\ \frac{4}{31} \\ \frac{5}{31} \\ \frac{15}{31} \\ \frac{1}{31} \end{bmatrix}. \tag{15}$$

Then, the vector Q_n of the probabilities of the next states in January 2020 is derived from (S. Q_c), while Q_a is the vector of the probabilities of the actual data set of states in January 2020.

$$Q_n = \begin{matrix} L \\ L^+ \\ M \\ M^+ \\ H \\ H^+ \end{matrix} \begin{bmatrix} \frac{28}{403} \\ \frac{229}{1612} \\ \frac{56}{403} \\ \frac{108}{403} \\ \frac{875}{1612} \\ 0 \end{bmatrix} \cong \begin{matrix} L \\ L^+ \\ M \\ M^+ \\ H \\ H^+ \end{matrix} \begin{bmatrix} 0.07 \\ 0.14 \\ 0.14 \\ 0.27 \\ 0.38 \\ 0 \end{bmatrix},$$

$$Q_a = \begin{matrix} L \\ L^+ \\ M \\ M^+ \\ H \\ H^+ \end{matrix} \begin{bmatrix} \frac{3}{31} \\ \frac{4}{31} \\ \frac{5}{31} \\ \frac{8}{31} \\ \frac{11}{31} \\ 0 \end{bmatrix} \cong \begin{matrix} L \\ L^+ \\ M \\ M^+ \\ H \\ H^+ \end{matrix} \begin{bmatrix} 0.10 \\ 0.13 \\ 0.16 \\ 0.26 \\ 0.35 \\ 0 \end{bmatrix}. \tag{16}$$

It is evident from this that the Markov model for the six states gives much less absolute errors than the Markov model for the three states. The absolute errors of the probabilities p (L), p (L+), p (M), p (M+), p (H), and p (H+) between 2019 and 2020 are 0.03, 0.01, 0.02, 0.01, 0.03, and 0, respectively. As a result, the CMMD method gave good results in forecasting by dividing the data set into an appropriate number of intervals. This can be applied to the types of data set that relates to various sciences.

4. Conclusion

This study discovered the forecasting of water quality by programming the regression models of the water elements in lakes. It is beneficial for predicting the corresponding values of other locations that have not been estimated yet or are impossible to reach. This study will help researchers to uncover the critical areas by reducing the time and effort to frequently move to significant water bodies. Consequently, a new algorithm has been reached for predicting water quality by the regression programming of the water elements in various depths of the lake. On the contrary, Markov transition probability matrix produces accurately forecasting for the different state probabilities of solar energy, whereby the performance of environmental energetic systems depend on solar radiation. In this study, the prediction method depends on the conversion of the data set into Markov models. This procedure to minimize the absolute errors of forecasting various states of solar energy in the future.

4.1. Future Work. This paper suggests the use of two prediction methods: algorithm of estimation regression models (AERMs) and converting data set to Markov model (CDMM) in the field of electronic industries such as lithium battery industry to increase its lifetime, reliability, and reduce failure rate. Moreover, the prediction of the system states to control commands and optimize the stochastic control strategy to achieve the speed control of the "model-free" quadrotor. It is worth mentioning that the two prediction methods reduce the time, effort, and cost in dealing with two types of data: the first type is limited and has high dispersion; in this case, the AERM algorithm will be used. When the second type of data are unlimited or large in size and need to be converted into a limited number of states, then CDMM method will be used. There is also another new proposal when there are two different items of the same system. Thus, it is required to predict the states of the system and items at the same time. For this reason, hidden Markov model will be used instead of Markov model. For example, a system contains two items: the virtual machines and their corresponding jobs. Therefore, it needs to predict the probability of the states of virtual machines (are they all in the busy states or not?) and the corresponding jobs (are they all in the processing states or not?). Consequently, these states will be predicted using the hidden Markov model.

Data Availability

(1) Physio-chemical properties and heavy metal concentrations of Cu, Cd, Pb, Zn, and Fe in various depths ranging from 0.13 to 2.5 m during January to March 2016 at 12 sampling sites that cover the northern part of Lake Manzala at the fish farming area called El Mussallas of Lake Manzala in Egypt. The data used to support the findings of this study are included within the article (Figure 1 in page 5) and (Table 1 in page 7). (2) Monthly mean daily global solar exposure in Queensland (Terrey Hills), Australia, from 1990 to 2021 data used to support the findings of this study are included within the supplementary information file named [IDCJAC0003_052081_Solar Radiation from 1990 to 2021_Queensland_Australia]. (3) Daily global solar exposure in Queensland (Terrey Hills), Australia, in 2019, data used to support the findings of this study are included within the supplementary information files named [IDCJAC0016_052081_2019_Solar Radiation_Queensland_Australia]. (4) Daily global solar exposure in Queensland (Terrey

Hills), Australia, in 2020, data used to support the findings of this study are included within the supplementary information file named [IDCJAC0016_052081_2020_Solar Radiation_Queensland_Australia]. (5) Daily global solar exposure in Queensland (Terrey Hills), Australia, in 2021, data used to support the findings of this study are included within the supplementary information file named [IDCJAC0016_052081_2021_Solar Radiation_Queensland_Australia].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the staff members of the Australian Bureau of Meteorology for providing climate data about Queensland, Australia, online. They also thank all contributing authors in this field of study.

Supplementary Materials

(1) The file named [Monthly Mean Solar Radiation] represents the monthly mean daily global solar exposure (MJ/m^2) in Queensland (Terrey Hills), Australia, from 1990 to 2021. (2) The file named [Solar Radiation 2019] shows the daily global solar exposure (MJ/m^2) in Queensland (Terrey Hills), Australia, in 2019. (3) The file named [Solar Radiation 2020] represents the daily global solar exposure (MJ/m^2) in Queensland (Terrey Hills), Australia, in 2020. (4) The file named [Solar Radiation 2021] shows the daily global solar exposure (MJ/m^2) in Queensland (Terrey Hills), Australia, in 2021; (*Supplementary Materials*)

References

[1] Z. Wu, X. Wang, Y. Chen, Y. Cai, and J. Deng, "Assessing river water quality using water quality index in Lake Taihu Basin, China," *The Science of the Total Environment*, vol. 612, pp. 914–922, 2018.

[2] Y. Qin, A. U. Alam, S. Pan et al., "Integrated water quality monitoring system with pH, free chlorine, and temperature sensors," *Sensors and Actuators B: Chemical*, vol. 255, no. 1, pp. 781–790, 2018.

[3] A. A. Elkady, S. T. Sweet, T. L. Wade, and A. G. Klein, "Distribution and assessment of heavy metals in the aquatic environment of Lake Manzala, Egypt," *Ecological Indicators*, vol. 58, pp. 445–457, 2015.

[4] M. Ali, "Assessment of some water quality characteristics and determination of some heavy metals in Lake Manzala, Egypt," *Egyptian Journal of Aquatic Biology and Fisheries*, vol. 12, no. 2, pp. 133–154, 2008.

[5] A. F. Abukila, "Assessing the drain estuaries' water quality in response to pollution abatement," *Water Science*, vol. 29, no. 1, pp. 1–18, 2015.

[6] M. Khadr and M. Elshemy, "Data-driven modeling for water quality prediction case study: the drains system associated with Manzala Lake, Egypt," *Ain Shams Engineering Journal*, vol. 8, no. 4, pp. 549–557, 2017.

[7] L. Jiang, Y. Li, X. Zhao et al., "Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China," *Ecological Modelling*, vol. 375, pp. 1–12, 2018.

[8] S. Sousa, F. Martins, M. Alvimferraz, and M. Pereira, "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations," *Environmental Modelling and Software*, vol. 22, no. 1, pp. 97–103, 2007.

[9] R. Aguiar and M. P. Collares, "Statistical properties of hourly global radiation," *Solar Energy*, vol. 48, no. 3, pp. 157–167, 1992.

[10] S. Bhardwaj, V. Sharma, S. Srivastava, O. S. Sastry, B. Bandyopadhyay, and S. S. Chandel, "Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model," *Solar Energy*, vol. 93, pp. 43–54, 2013.

[11] F. Besharat, A. A. Dehghan, and A. R. Faghih, "Empirical models for estimating global solar radiation: a review and case study," *Renewable and Sustainable Energy Reviews*, vol. 21, pp. 798–821, 2013.

[12] K. Yadav and S. S. Chande, "Solar radiation prediction using Artificial Neural Network techniques: a review," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 772–781, 2014.

[13] M. Ali, D. Khan, M. Amir, A. Ali, and Z. Ahmad, "Predicting the direction movement of financial time series using artificial neural network and support vector machine," *Complexity*, vol. 2021, Article ID 2906463, 13 pages, 2021.

[14] R. Niaz, M. Almazah, X. Zhang, I. Hussain, and M. Faisal, "Prediction for various drought classes using spatiotemporal categorical sequences," *Comlexity*, vol. 2021, Article ID 7145168, 11 pages, 2021.

[15] R. Niaz, X. Zhang, N. Iqbal, M. Almazah, T. Hussain, and I. Hussain, "Logistic regression analysis for spatial pattern of drought persistence," *Complexity*, vol. 2021, Article ID 3724919, 13 pages, 2021.

[16] K. Liu, X. Hu, J. Meng, J. M. Guerrero, and R. Teodorescu, "RUBoost-based ensemble machine learning for electrode quality classification in Li-ion battery manufacturing," *IEEE/ASME Transactions on Mechatronics*, 2021.

[17] K. Liu, X. Hu, H. Zhou, L. Tong, D. Widanalage, and J. Marco, "Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 2944–2955, 2021.

[18] H. Hu and Q. L Wang, "Proximal policy optimization with an integral compensator for quadrotor control," *Frontiers of Information Technology and Electronic Engineering*, vol. 21, pp. 777–795, 2020.

[19] Q. Wang, H. E. Psillakis, C. Sun, and F. L. Lewis, "Adaptive NN distributed control for time-varying networks of nonlinear agents with antagonistic interactions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2573–2583.

[20] M. S. Beheary and F. A. El-Matary, "Risk evaluation of heavy metal in sediments of the fish farming area in the Mediterranean section of Lake Manzala," *Scientific Journal for Damietta Faculty of Science*, vol. 1, no. 4, pp. 69–78, 2015.

[21] G. A. H. Sallam and E. A. Elsayed, "Estimating relations between temperature, relative humidity as independed variables and selected water quality parameters in Lake Manzala, Egypt," *Ain Shams Engineering Journal*, vol. 9, pp. 1–14, 2018.

[22] O. Barakat, A. Mostafa, T. L. Wade, S. T. Sweet, and N. B. El Sayed, "Assessment of persistent organochlorine pollutants in

sediments from Lake Manzala, Egypt," *Marine Pollution Bulletin*, vol. 64, pp. 1713–1720, 2012.

[23] M. M. El Genidy, "Multiple nonlinear regression model for the maximum number of migratory bird types during migration years," *Communications in Statistics-Theory and Methods*, vol. 46, no. 16, pp. 7969–7975, 2017.

[24] B. Gong and J. Ordieres-Mere, "Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong," *Environmental Modelling and Software*, vol. 84, pp. 290–303, 2016.

[25] M. S. Mecibah, T. E. Boukelia, R. Tahtah, and K. Gairaa, "Introducing the best model for estimation the monthly mean daily global solar radiation on a horizontal surface (case study: Algeria)," *Renewable and Sustainable Energy Reviews*, vol. 36, pp. 194–202, 2014.

[26] M. M. El Genidy, "Multiple nonlinear regression of the Markovian arrival process for estimating the daily global solar radiation," *Communications in Statistics-Theory and Methods*, vol. 48, no. 22, pp. 5427–5444, 2019.

[27] S. Park, M. S. Gil, H. Im, and Y. S. Moon, "Measurement noise recommendation for efficient Kalman filtering over a large amount of sensor data," *Sensors*, vol. 19, 2019.