Journal of Modeling and Optimization 2020;12(1):23-37
https://doi.org/10.32732/jmo.2020.12.1.23

Pre-processing Protocol for Nonlinear Regression of Uneven Spaced-Data

Palash Panja[1,2,*], Pranay Asai[2], Raul Velasco[1], Milind Deo[2]

1. Energy & Geoscience Institute, University of Utah, Salt Lake City, Utah, USA
2. Department of Chemical Engineering, University of Utah, Salt Lake City, Utah, USA
Email: ppanja@egi.utah.edu (Corresponding author)

Received: 4 February 2020; Accepted: 5 March 2020; Available online: 15 April 2020

Abstract: Regression of experimental or simulated data has important implications in sensitivity studies, uncertainty analysis, and prediction accuracy. The fitness of a model is highly dependent on the number of data points and the locations of the chosen points on the curve. The objective of the research is to find the best scheme for a nonlinear regression model using a fraction of total data points without losing any features or trends in the data. Six different schemes are developed by setting criteria such as equal spacing along axes, equal distance between two consecutive points, constraint in the angle of curvature, etc. A workflow is provided to summarize the entire protocol of data preprocessing, training and testing nonlinear regression models with various schemes using a simulated temperature profile from an enhanced geothermal system. It is shown that only 5% of data points are sufficient to represent the entire curve using a regression model with a proper scheme.

Keywords: Nonlinear regression; Data pre-processing; Time series; Uneven interval; Data reduction.

1. Introduction

Data extraction, processing and interpretation have become pivotal tool in making informed and risk evaluated decision in every industry. Time series data is widely used to forecast for weather, disease outbreak, stock, production and many more[1-18].

In absence of process/field data, simulation is a systematic alternative tool to generate data by mimicking physical system by solving complex set of equations. However, the data generated by simulation is in a very raw/crude form and could not be directly used to develop predictive models. One of the major problems faced is that the data points are not evenly distributed over a time period. This is caused because of the different convergence techniques used by most numerical simulators. Each simulator has a pre-defined convergence limit which is guided by the minimum time-step provided to the simulator. The minimum time-step is defined to make sure the equations converge and doesn't give any errors or doesn't introduce any artifacts in the results. The initial time-step is chosen carefully according to the nature of equations used and depending on the physical process and the time scale. For example, a simulation with a fixed time-step would generate large number of data points and would also require more time to run. Whereas, using adaptive time-step, the simulator initially generates data points at very small-time interval and as the equations begin to converge, it gradually increases the time-step and hence increasing the interval for data point generation. This leads to comparatively small number of data points as compared to the fixed time-step but even then, the result might contain unnecessary amount of data points.

In this study, we have considered data obtained from a commercial simulator for a temperature decline curve in an enhanced geothermal system [19]. Other researchers [20-22] also studied the temperature profile in geothermal system by simulating complex system or solving simpler system analytically. As described before that uneven intervals in the data points are common in simulated results especially in time series data. To demonstrate this numerical fact, produced water temperature from our previous study [19] which consists of 9936 points from an enhanced geothermal system is shown in Figure 1.

The logarithmic scale is used to expand the early time. The actual shape of the temperature profile in linear scale is also shown inside the figure. To investigate the unevenness of data points, the X-axis i.e., time is divided into five groups such a way that each group contains 20% data. The first 20% of the data is from 0-23 days, whereas last 20% of the data is from 6496 days (from 3504 to 10000 days). It is evident from this distribution that the data is highly dense towards initial time period. The density of data (number of points per unit time) is reduced towards the terminal time. In some instances, localized dense data is also observed due to curvature i.e., changes in slope along curve. More data points are required to represent any curvature compared to fixed slope or straight-line portion of the curve.

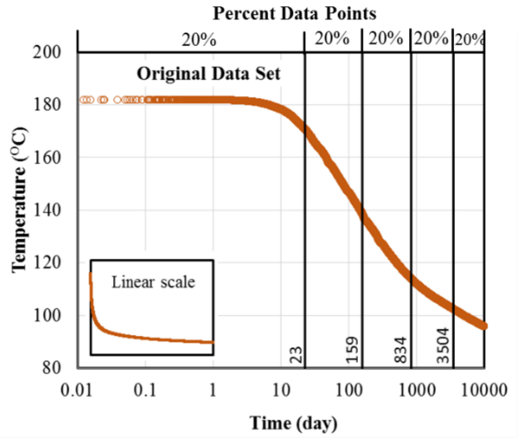


Figure 1. The data density along the X-axis

This uneven distribution of data points over the entire time period poses a problem in regression of a nonlinear mathematical equation because the data points are drastically skewed towards the beginning of the simulation. Thus, deeming the data points towards the end of simulation as the outliers and hence sometimes model cannot be fitted properly. It is not always practical to choose all points for regression of a model because it requires computer memory and power and it doesn't always guarantee a better fit.

The study focuses on tackling such uneven distribution in the simulation results by implementing smart and novel techniques on preprocessing of the data so that it could be used in regression or curve fitting. To validate proposed data processing techniques, we considered the simulated data for the temperature decline curve in an enhanced geothermal reservoir generated through a commercial simulator.

The objective of this study is to reduce the total number of points to represent the entire curve to facilitate post processing of data such as curve fitting. Curve fitting using regression is highly dependent on number points as well as the local density of points. Various schemes are investigated to reduce the total number points and to obtain a better fit.

Usually while performing curve fitting in linear models the coefficient of determination, R^2 value is considered as the benchmark to establish the fitness of the curve, with R^2 approaching 1 being the best. However, this is not applicable for nonlinear curve fitting. Spiess and Neumeyer [23] shows how R^2 is an inadequate measure to validate the fitness of curve in nonlinear models. We have used error, R^2 and normalized root mean square error (NRMSE) to rank different schemes.

2. Methodology

Various steps involved in the proposed protocol to reduce number of points in an uneven time series for regression are discussed here. All steps are summarized in a workflow for better comprehension in Figure 2. The workflow can be divided into three broad sections namely data preprocessing, training of mathematical model and testing of the fitted model.

Few individual components (bold in the figure 2) of the workflow such as normalization of data, followed by selection of schemes and mathematical model for curve fitting are discussed in details.

2.1 Data normalization

It is observed in the most of cases if not all that the ranges (minimum to maximum) and actual values of dependent (Y-axis) and independent (X-axis) variables are not comparable in the same scale. For example, in figure 1, the temperature (Y-axis) varies from 92 to 182 ^{O}C (range 90 ^{O}C) whereas time (X-axis) varies from 0 day to 10,000 days (range 10,000 days). Therefore, time is more sensitive to regression compared to temperature. To avoid this mathematical problem, it is advised to normalize the data to 0 to 1 for both axes ensuring the same ranges. In the context of a geothermal system, the temperature and time are normalized as shown in Equations 1 and 2.

$$\bar{T} = \frac{T-T_{min}}{T_{max}-T_{min}} \tag{1}$$

$$\bar{t} = \frac{t-t_{min}}{t_{max}-t_{min}} \tag{2}$$

Using the above formulae, the minimum temperature i.e., 92 ^{O}C becomes zero and maximum temperature 182 ^{O}C becomes 1. Similarly, minimum and maximum times become 0 and 1 too.

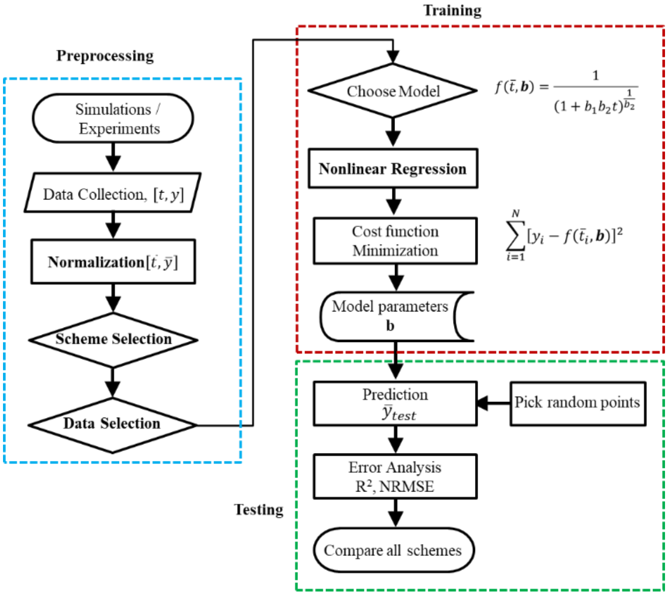

Preprocessing
Simulations / Experiments
Data Collection, $[t, y]$
Normalization$[\bar{t}, \bar{y}]$
Scheme Selection
Data Selection
Training
Choose Model
$f(\bar{t}, \boldsymbol{b}) = \frac{1}{(1 + b_1 b_2 t)^{\frac{1}{b_2}}}$
Nonlinear Regression
Cost function Minimization
$\sum_{i=1}^{N} [y_i - f(\bar{t}_i, \boldsymbol{b})]^2$
Model parameters b
Testing
Prediction $\bar{y}_{test}$
Pick random points
Error Analysis R^2, NRMSE
Compare all schemes

Figure 2. Workflow for data preprocessing, training and testing of fitted model

2.2 Scheme selection

Using the normalized data, six schemes are investigated in this study to select certain number of points from total points of 9936 (see figure 1) as shown in the Table 1.

Table 1. Various schemes for selecting points to reduce the number of points

Label	Method
Scheme 1	Entire Original Data set
Scheme 2	Equal division of X-axis
Scheme 3	Equal division of Y-axis
Scheme 4	Equal division along curve
Scheme 5	Constraint in deflection
Scheme 6	Mixed of schemes 3 and 4

2.2.1 Scheme 1: Entire data set

In this scheme, data set is kept unchanged i.e., entire data of 9939 points are chosen for regression. This is base case for comparison with other schemes to establish the effectiveness of the data preprocessing. All points are plotted in Figure 3.

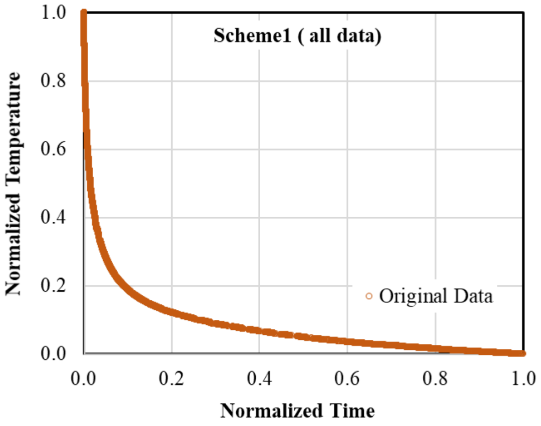


Figure 3. Scheme 1 where all data points are used for regression

2.2.2 Scheme 2: Equal division of X-axis

In this scheme, the temperatures are selected with equal interval of time. Spacing between two consecutive points in this scheme is calculated as given in equation 3.

$$\Delta X = \Delta t = \frac{t_{max} - t_{min}}{N-1} = \frac{t_{max}}{N-1}, \; where \;\; t_{min} = 0 \tag{3}$$

The number of data points, N, is decided based on the sensitivity studies. Time for any location can be calculated by equation 4.

$$t_i = t_{min} + \Delta T\,(i-1), \;\; i = 1,2,3..N \tag{4}$$

After choosing the value in the X-axes, the temperature is calculated based on linear interpolation of two nearest points. Details discussion of sensitivity of number of points on curve fitting is provided later. 1% data (100 points from 9939 points) is considered as case 1 in the sensitivity study and it is used for all figures showing different schemes for demonstration. Distribution of points for scheme 1 is shown in Figure 4.

It is evident from the figure 4 that the distance between two consecutive points varies depending on the curvature or slope of the curve with respect to time ($\Delta T/\Delta t$). More points are located at lower slope section such as the flat potion of the curve compared to higher slope section Technically point density can be defined as the number of points in a fixed one-dimensional length (curve or straight line). Same point density at any location on the curve can be found for the curve with constant slope i.e., for linear equation.

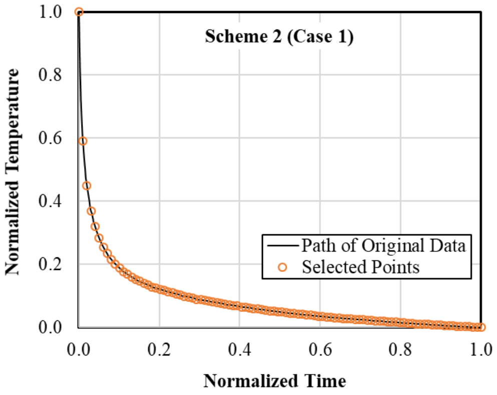


Figure 4. Scheme 2 where points are located with equal interval in X-axis i.e., time

2.2.3 Scheme 3: Equal division of Y-axis

Like the equal division of X-axis, the entire range of Y-axis can be divided into equal interval as shown in Figure 5.

Spacing in this scheme is calculated as given in equation 5.

$$\Delta Y = \Delta T = \frac{T_{max} - T_{min}}{N-1} \tag{5}$$

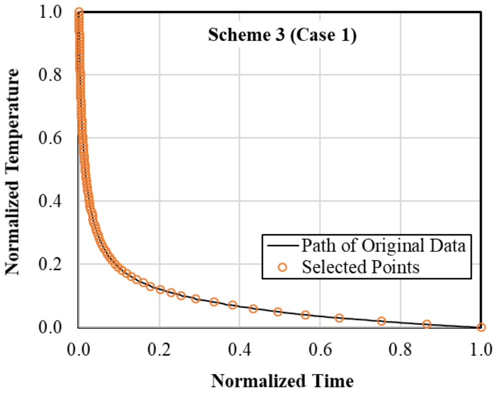


Figure 5. Scheme 3 where points are located with equal interval in Y-axis i.e., temperature

In the case of a geothermal system, temperature starts initially at maximum value and reduces towards the minimum as time goes. Any data point in this scheme is calculated by equation 6.

$$T_i = T_{max} - \Delta T\ (i-1),\ \ i = 1,2,3..N \tag{6}$$

Location of higher point density in the curve is opposite to the previous scheme. A higher point density is found in the higher slope section.

2.2.4 Scheme 4: Equal division along curve

In this scheme, the entire curve is divided into equal pieces along the trajectory of the curve. First task in this scheme is to calculate the length of the entire curve. Next, the total length is divided into certain number with equal spacing along the curve. The distance between two consecutive points (ΔL) is calculated using equation 7 and demonstrated in Figure 6.

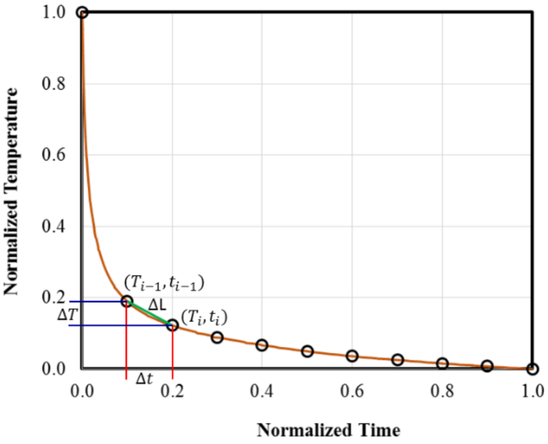


Figure 6. The demonstration of intervals and distance between two consecutive points

$$\Delta L_{i,i-1} = \sqrt{(T_i - T_{i-1})^2 + (t_i - t_{i-1})^2} \tag{7}$$

To calculate the length of the entire curve, distances between two consecutive points (as calculated using equation 7) are added together as shown in equation 8.

$$L = \sum_{i=2}^{N} \Delta L_{i,i-1} = \sum_{i=2}^{N} \sqrt{(T_i - T_{i-1})^2 + (t_i - t_{i-1})^2} \tag{8}$$

If the entire length along the curve is divided into N equal spacing, then the interval is calculated as

$$\Delta L = \frac{L}{N-1} \tag{9}$$

Placing points with ΔL interval along curve is calculated as

$$t_i = t_{i-1} + \sqrt{\Delta L^2 - (T(t_i) - T_{i-1})^2} \tag{10}$$

As shown in equation 10, the method to find out the points with ΔL interval with previous point is an iterative method. The current time step (t_i) is assumed first, the temperature (T_i) is interpolated from the original data set. Then, using calculation shown in equation 10, current time step (t_i) is calculated again. If the assumed and calculated time match, then current time step is accepted and proceed for next time step. The points calculated in this way are shown in Figure 7.

In this scheme, it is clear that each section of the curve has same point density irrespective of the slope of the curve.

2.2.5 Scheme 5: Constraint in deflection

In the previous schemes, especially in the schemes 2 to 4, the fixed number of total points is selected based on their criteria. Selection of total number of points is totally knowledge based. The total number of points should be sufficient to capture all the features of the curvature. Scheme 5 is formulated to ensure that the curved sections with sharp changes in slope get sufficient points during regression. In this scheme, sum of change in slopes (in terms of angle) in successive points or deflection in the curve is set as a criterion. The calculation of angle in this scheme is explained in Figure 8.

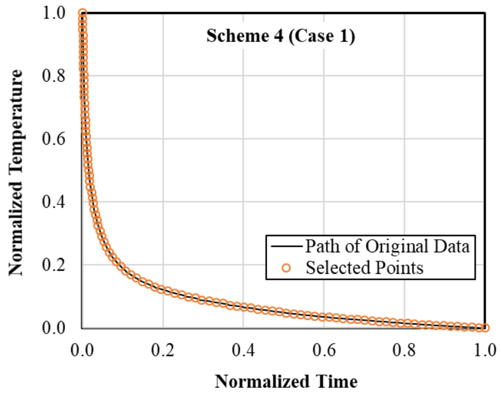


Figure 7. Scheme 4 where points are chosen based on equal division of the length of the curve

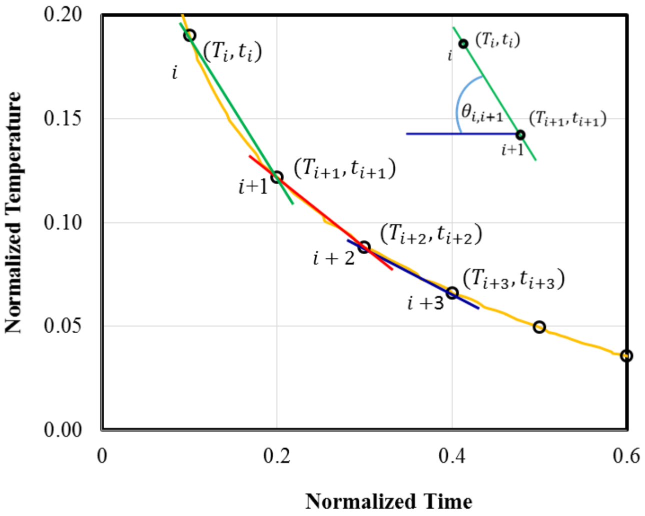


Figure 8. Deflection based scheme where points are chosen based on user specified angle of deflection

The slope and corresponding angle between two consecutive points are calculated as described in equations 11 and 12.

$$S_{i,i+1} = \left(\frac{T_{i+1}-T_i}{t_{i+1}-t_i}\right) \tag{11}$$

$$\theta_{i,i+1} = \tan^{-1}(S_{i,i+1}) \tag{12}$$

To calculate the total deflection from the (i+1)th point, angles between next few points such as $\theta_{I,i+1}$, $\theta_{i+1, i+2}$, $\theta_{i+2, i+3}$ should be known. The total change in angle from point I to next k points is calculated as

$$\Delta\theta_{i,k} = (\theta_{i+1,i+2} - \theta_{i,i+1}) + (\theta_{i+2,i+3} - \theta_{i+1,i+2}) + \cdots + (\theta_{i+k,i+k+1} - \theta_{i+k-,i+k})$$
$$= \sum_{j=1}^{k}(\theta_{i+j,i+j+1} - \theta_{i+j-1,i+j}) \tag{13}$$

Then the percentage change in slope with respect to angle between I and i+1 points is calculated as

$$\Delta\theta_i = \left|\frac{\theta_{i+1,i+2} - \theta_{i,i+1}}{\Delta\theta_{i,k}}\right| \text{x } 100 \tag{14}$$

Points are added until the calculated $\Delta\theta_I$ reaches the set criteria of deflection. If the calculated $\Delta\theta_I$ according to equation 14 is acceptable for a given value (say 75%), then the kth point after ith point is acceptable as the next i+1 point in the scheme. The points calculated this way with 75% of deflection tolerance are shown in Figure 9.

This scheme ensures that no parts of curves with sharp slope change are ignored for the regression. Intervals in the curve are irregular depending on the localized slope. In original data, sudden changes of slope are observed in many places and it causes some close spacing of points. This cause the different weightage of different section of the curve in the cost function.

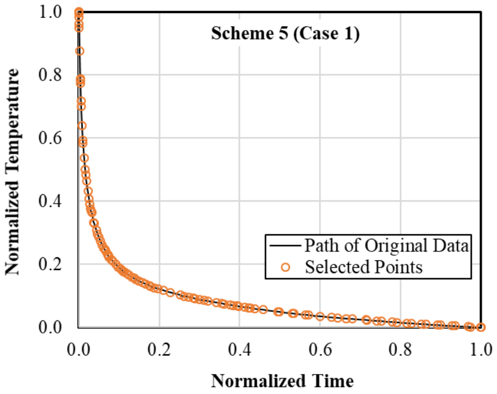


Figure 9. Scheme 5 where deflection is chosen as criteria to select representative points

2.2.6 Scheme 6: Mix of schemes 4 and 5

To reduce the number of total points further, the scheme 4 and 5 are mixed together. The scheme 6 is same as scheme 5 except that the points generated in scheme 4 are considered instead of original data set (scheme 1). In scheme 4, equally spaced points along curves are generated. A few neighboring points might not have sharp changes in slopes. Using this scheme those neighboring point could be merged together. The points are significantly reduced compared to figures 7 and 8 as shown in Figure 10.

It is noticed that four points out of seven points are located on the deflection section. On the other hand, few points are located on the straight-line sections.

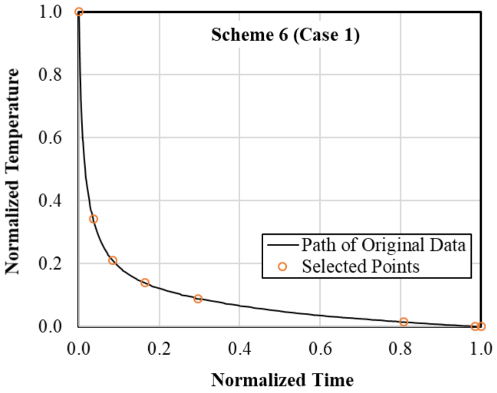


Figure 10. Selection of points based on 75% cumulative deflection change on the equally spacing points along curve

2.3 Data selection

Total 9939 points are collected to represent 0 to 10,000 days temperature versus time profile from simulated enhanced geothermal system [19]. To study the sensitivity of the chosen number of data points on the fitted curve, 8 cases (from 1 to 20 % of total data) are investigated as given in Table 2.

In scheme 1, 100% data is used, therefore only single case is applied for scheme 1 and not shown in the table 2. For schemes 2 to 5, 1 to 20% of total data are used to estimate the model parameters in equation 15. In scheme 6, number of data points (0.08% to 6.9%) are less than the data points used in scheme 4 because of the nature of the scheme. In scheme 5, choosing any number of points is not straightforward like scheme 2 to 4. In this scheme, the constraint in deflection (angle of curvature) is set as criteria instead of number of points. Therefore, it is a trial and error method to choose an angle such a way that the scheme will have certain number of points. Angles for scheme 5 are 75, 15.5, 2.6, 1.07, 0.70, 0.235, 0.127 and 0.0785 for cases 1 to 8 respectively.

Table 2. Data utilization in each scheme for various cases. (100% data is used for scheme 1)

Case	Percentage of total data		Number of data	
	Scheme 2 to 5	Scheme 6	Scheme 2 to 5	Scheme 6
Case 1	1	0.08	100	8
Case 2	2	0.3	199	30
Case 3	3	1.1	299	103
Case 4	4	1. 8	398	177
Case 5	5	2.2	497	214
Case 6	10	3.8	994	375
Case 7	15	5.4	1491	533
Case 8	20	6.9	1988	690

2.4 Nonlinear regression

After choosing points from simulated data based on the criteria set by each scheme, a mathematical model is fitted. Multivariable regression model is proposed by researchers [24] where the parameters varies with time. Another model namely local linear regression model [25] is also evaluated using limited amount of data unlike ARIMA. Reviewing several time series forecasting models which are available now, De Gooijer and Hyndman [26] classified them into eight categories namely (i) exponential smoothing[27-29], (ii) Autoregressive Integrated Moving Average (ARIMA) [30, 31], (iii) seasonal models[12, 13], (iv) state space and structural models and the Kalman filter [32-34], (v) nonlinear models [35, 36], (vi) long-range dependence models, e.g. the family of Autoregressive Fractionally Integrated Moving Average (ARFIMA) models[15, 37], (vii) Autoregressive Conditional Heteroscedastic/Generalized Autoregressive Conditional Heteroscedastic (ARCH/GARCH) models [16, 38, 39] and (viii) count data forecasting[11, 18]. In this study, a nonlinear function, $f(\bar{t},\boldsymbol{b})$ is chosen for regression as shown in Equation 15.

$$\bar{T} = f(\bar{t},\boldsymbol{b}) = \frac{\bar{T}_0}{(1+b_1 b_2 t)^{\frac{1}{b_2}}} \tag{15}$$

$$where,\ \bar{T}_0 = \frac{T(t_{min})-T_{min}}{T_{max}-T_{min}} \tag{16}$$

The above equation is originally applied by Arps [40] for decline in oil rate. In case of normalized data (0 to 1), the $(\bar{T}_0)$ becomes one, therefore only parameters required to determine using regression are b_1 and b_2. In the curve fitting method, a cost function which is generally the sum of the errors or square of the sum of errors is minimized. An optimization routine '*nlinfit*' in Matlab (Mathworks Inc.) is used to minimize the cost function as given in Equation 17.

$$cost\ function = \sum_{i=1}^{N}[y_i - f(\bar{t}_i,\boldsymbol{b})]^2 \tag{17}$$

The y_i is the normalized temperature from the experiments or simulations and $f(\bar{t},\boldsymbol{b})$ is the normalized temperature predicted by the model for same normalized time. Total N points or observations are used for the fitting. The quality of the fitted function is evaluated by various statistical measurements such as coefficient of determination (R^2) and normalized root mean square error (NRMSE) which are discussed in the Appendix 2.

3. Results and discussions

As described in the workflow (figure 2), the results are analyzed for both training and test data sets. Temperature profiles and errors in predictability are discussed here for scheme 1 to 7 for various scenarios.

3.1 Training of models

Model provided in equation 15 is trained for all the cases of various schemes. Model parameters b_1 and b_2 for all schemes are enlisted in Appendix 1. It is evident that the only minor variations in parameters (b_1 and b_2) for different cases are observed. It implies that the parameters are independent of the number of points during the training. The fitted model from various scheme and the original observation are compared in Figure 11 for case 1 (1% data). Other cases are provided in Appendix 1. Although the scheme 1 doesn't fall under any cases (case 1 to case 8), it is displayed with all the figures for comparisons as base case.

Although the values of model parameters, b1 and b2 (see Table A1.1) vary from scheme to scheme, a good agreement is noticed between fitted model and simulated data. This indicates that the combination of values b_1 and

b_2 in the model (equation 15) works well in prediction. Error is plotted in Figure 12 to visualize the difference between fitted model and simulated data.

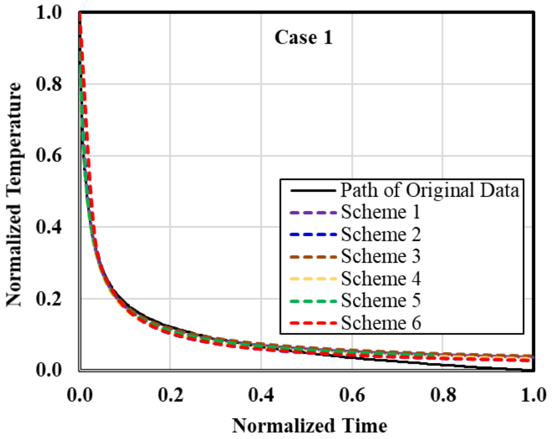


Figure 11. Comparison of fitted functions using schemes 1 to 6 for case 1 with original data

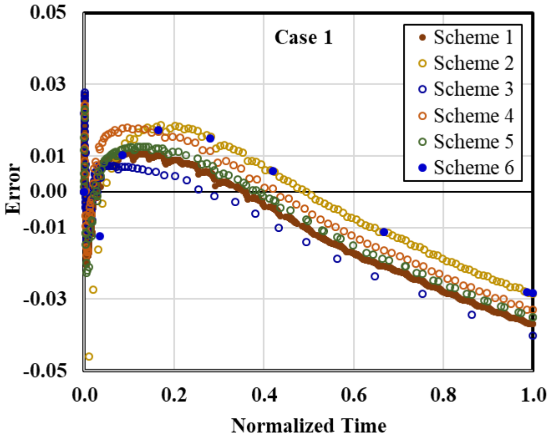


Figure 12. Errors using case 1 for schemes 1 to 6

Because of the fact that data was normalized, all points fall between 0 and 1, the errors in the range of -0.05 to 0.05 are insignificant. Errors vary with time for all schemes. The highest errors are observed at the end of the curve and at the point of deflection of the curve. It implies that the fitted model had difficulty to represent the curvature section of the data. Scheme 3 had low errors initially but it started deviating in the later time. Surprisingly, scheme 1 where 100% data are used had significant errors. Schemes 2, 4, 5 and 6 showed better fit compared to other schemes by showing overall less error for entire time period. Although, scheme 2 had highest error in the initial portion. Schemes 4 and 5 are possibly the best fit where low errors are observed in the curvature section. This is because of the nature of the selection criteria of points in schemes 4 and 5. All points are equally spaced on the curve providing equal weightage to each section of curve in regression. On the other hand, in scheme 5, the curve section had sufficient point for more weightage in the curve fitting, in other words, the cost function defined in equation 17 is more influenced by this section.

Overall fitness of the model for various cases in different schemes is measured by the combined error such as coefficient of determination (R^2) and normalized root mean square error (NRMSE) as shown in Figure 13.

Different percentage in the X-axis in the above figures are the different cases as shown in table 2. Because scheme 1 has only one case (100% data), one R^2 and one NRMSE values are calculated which are shown by dotted blue lines in the figures. Although the coefficient of determinations (R^2) of fitted curves for all schemes except scheme 2 in figure 13(a) are high which an indication of good fit, the NRMSEs are also in the higher side for schemes 4 and 6 which is an indication of bad fitting.

Considering the results from figures 12 and 13, schemes 4 and 5 are the best choice to select points for regression. Scheme 2 could be another choice but the initial deviations (see figure 12) make it unsuitable. Another observation is that the NRMSE values decrease until 10% data in the most of the schemes. Therefore, 5-10% data can be considered as optimum for regression of all schemes. In the next section of testing model, we continue with model parameters which are obtained from 5% data set (case 5).

3.2 Testing of models

Fitness of a model is not always guaranteed by the error analysis from training data. Test data set which is essentially randomly picked points within the range of study is required to check the predictability of the model. In testing of various schemes, model parameters (b1 and b2) obtained from case 5 are chosen. As shown in the workflow (figure 2), one hundred random values of time are chosen in the range of 0 to 1 for testing of the model. The predicted values from scheme 1 to scheme 6 are compared with the simulated data in Figure 14.

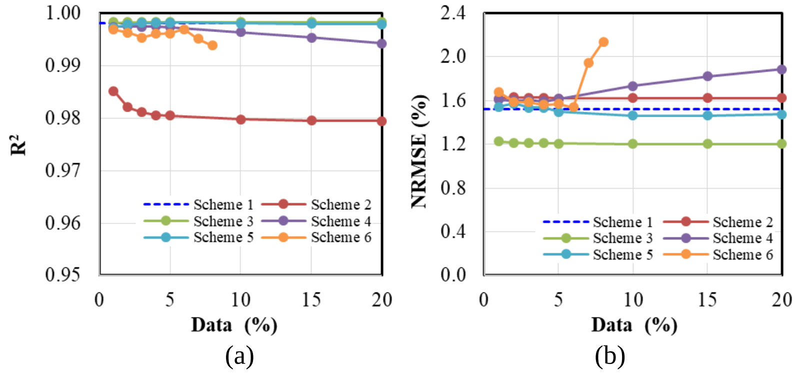


Figure 13. Error analysis of fitted functions for various scheme and different cases (a) Coefficient of determination (R^2) (b) Normalized Root Mean Square Error (NRMSE)

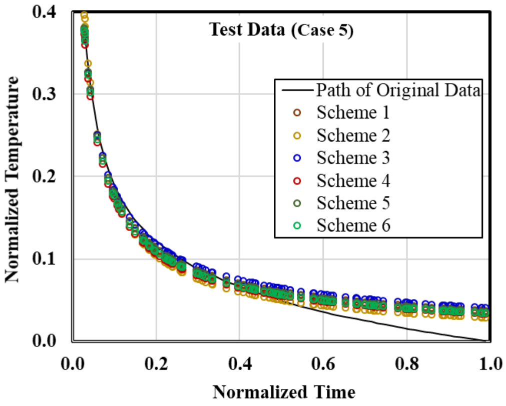


Figure 14. Testing of various scheme for model fitness using randomly chosen 100 values of time

The predicted values from different schemes have discernible differences. All schemes predicted well close to the observed values until the midway. Values are underestimated towards the end. To differentiate each scheme, the error plots are shown in Figure 15.

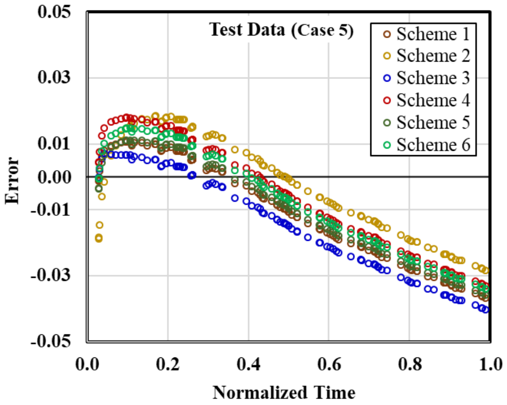


Figure 15. Errors in the predictions from scheme 1 to 6 for testing data

Like training set, scheme 2 has the highest error in the initial time period but errors are low in the later time. Schemes 4, 5 and 6 have the low variations in error throughout the entire time period. This can be evident from the quantitative error analysis (R^2 and NRMSE) as shown in Figure 16.

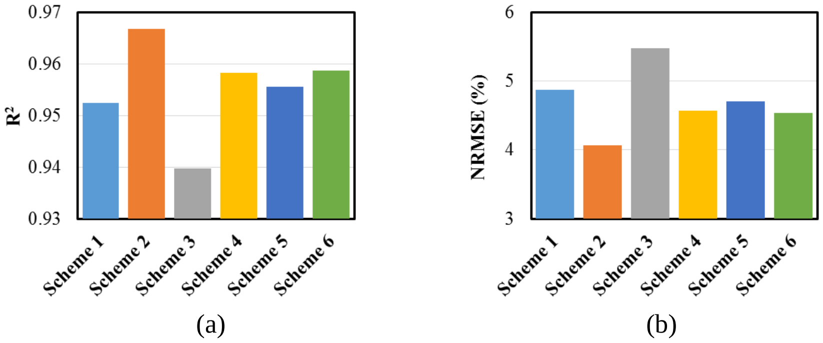


Figure 16. Quantitative error analysis of various scheme for testing set (a) Coefficient of determination (R2) (b) Normalized Root Mean Square Error (NRMSE)

The scheme 2 has the highest R^2 and the lowest NRMSE which indicate the best fit among all scheme, however, due the higher initial error (figure 16), this scheme may not be the best to predict for entire time period. Scheme 3 has the lowest R^2 and the highest NRMSE values which make it least suited scheme among all. Like training data set, schemes 4, 5 and 6 remain the best schemes where higher R^2 and lower NRMSE are observed.

4. Conclusions

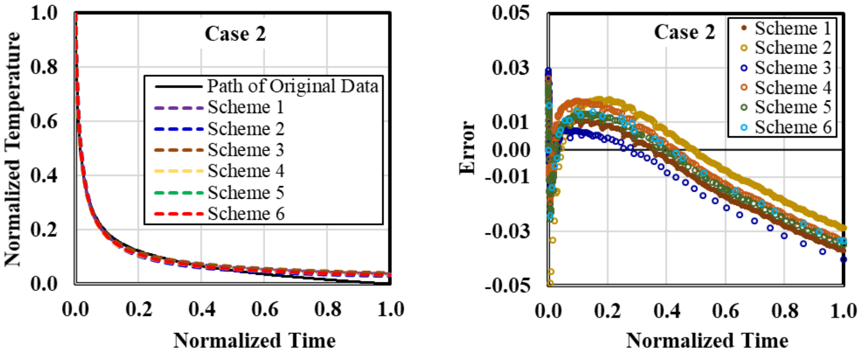
Despite the large availability of data points, fitted functions often fail to represent all these points due to differences in data density. 1% to 20% data points are selected using 6 different selecting criteria (schemes 1 to 6). It is shown that even 1% worth of data points is as good as the entire data set for regression as long as the proper scheme to select points from original data set is chosen. 5-10% data points can be considered as optimum value. Scheme 4 (where distance between two consecutive points is fixed) and 5 (where angle of deflection is the criteria) are the most efficient schemes that provide a better fit of the mathematical model for any given number of points. Test data set which is chosen randomly confirms the robustness of these schemes. This study helps reduce the number of data points necessary during regression and improve the fit of any model when the data points are not evenly distributed.

Appendix 1

Model parameters from different schemes are summarized in Table A1.1. Errors in different schemes for various cases are compared in Figures A1.1 to A1.7.

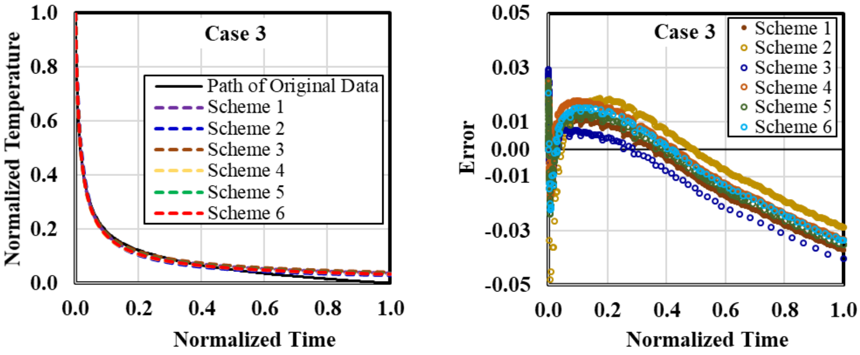
Table A1.1: Model parameters b_1, b_2 after regressions of models (equation 15) from various schemes for different cases

Model Parameter	Case	Scheme 1	Scheme 2	Scheme 3	Scheme 4	Scheme 5	Scheme 6
b1	Case 1	1.40	1.20	1.47	1.34	1.37	1.19
	Case 2		1.21	1.47	1.34	1.36	1.33
	Case 3		1.21	1.47	1.35	1.37	1.34
	Case 4		1.21	1.47	1.35	1.37	1.34
	Case 5		1.20	1.47	1.34	1.38	1.35
	Case 6		1.21	1.47	1.33	1.37	1.37
	Case 7		1.21	1.47	1.30	1.37	1.28
	Case 8		1.21	1.47	1.28	1.37	1.22
b2	Case 1	72.2	59.1	75.5	71.5	70.2	57.2
	Case 2		59.5	75.5	71.7	69.7	68.4
	Case 3		59.5	75.6	71.8	70.5	69.8
	Case 4		59.5	75.6	71.7	70.1	70.2
	Case 5		59.4	75.6	71.7	70.3	70.2
	Case 6		59.4	75.6	71.4	68.9	71.0
	Case 7		59.4	75.6	70.3	68.6	67.7
	Case 8		59.4	75.6	69.1	68.6	65.3



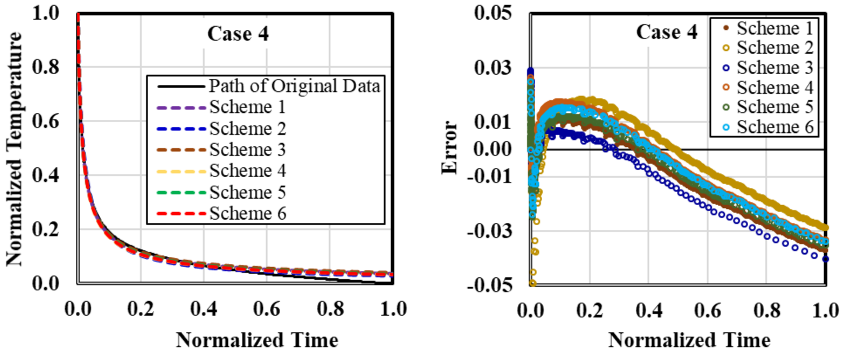
(a) (b)

Figure A1.1: Case 2 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data



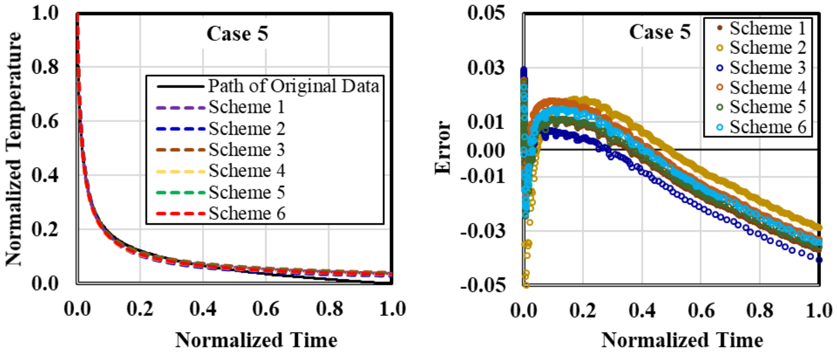
(a) (b)

Figure A1.2: Case 3 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data



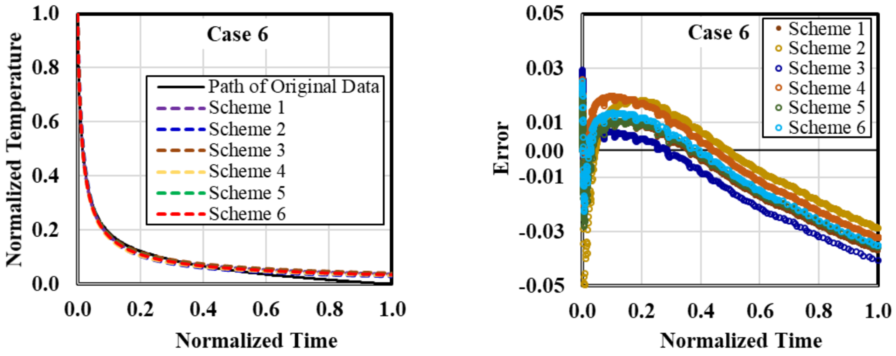
(a) (b)

Figure A1.3: Case 4 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data



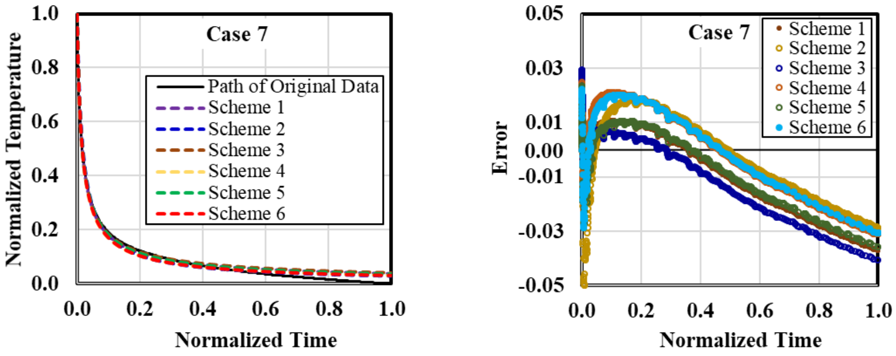
(a) (b)

Figure A1.4: Case 5 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data



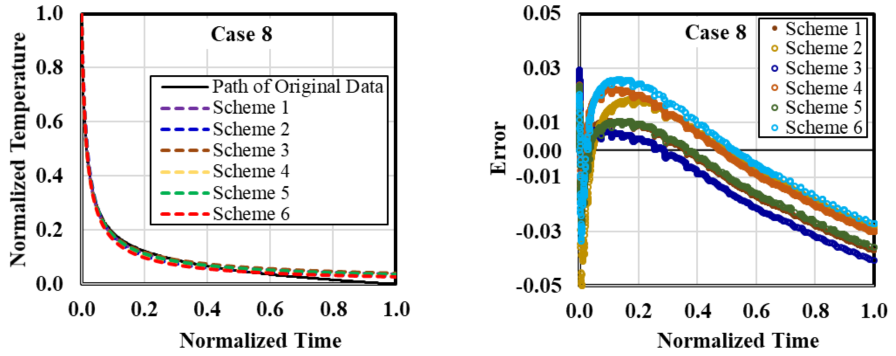
(a) (b)

Figure A1.5: Case 6 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data



(a) (b)

Figure A1.6: Case 7 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data



(a) (b)

Figure A1.7: Case 8 (a) comparison of predicted data from all schemes with original data (b) error in the predicted data

Appendix 2

1) The coefficient of determination (R^2):

The overall accuracy of a regression is measured by the coefficient of determination, R^2 which is defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{A.1}$$

where, $SS_{res} = \sum_{i=1}^{n}\left(Y_{obs,i} - Y_{model,i}\right)^2$ is the residual sum of squares, $SS_{tot} = \sum_{i=1}^{n}\left(\bar{Y}_{obs} - Y_{model,i}\right)^2$ is the total sum of squares and $\bar{Y}_{obs} = \frac{1}{n}\sum_{i=1}^{n} Y_{obs,i}$ is the mean of observed values.

The values of R^2 vary from 0 to 1. The R^2 value closed to one are indication of better fit of the model curve with observed data.

2) Normalized Root Mean Square Error (NRMSE):

The error in the fitted model is calculated by the difference between the actual or measured value and the predicted value by the model as given in Equation A.2

$$Error, e_i = Y_{obs,i} - Y_{model,i} \tag{A.2}$$

The error calculated from equation A.2 is for a single point. The total error for all points can be measured using mean square error in Equation A.3

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n} = \frac{\sum_{i=1}^{n}\left(Y_{obs,i} - Y_{model,i}\right)^2}{n} \tag{A.3}$$

The Root Mean Square Error (RMSE) (also known as the root mean square deviation, RMSD), is used to measure the cumulative error for the entire curve.

Often, square root of the MSE or the square root of the mean squared error (RMSE) is used to measure the fitness:

$$RMSE = \sqrt{MSE} \tag{A.4}$$

where Y_{obs} is observed values and Y_{model} is modeled values.

It is not always fair to analyze the error in terms of absolute values because different schemes may have different absolute values and their ranges. Non-dimensional form of the RMSE is used instead by normalizing RMSE with the range of the observed data to obtain Normalized Root Mean Square Error (NRMSE) as given in Equation A.5

$$NRMSE = \frac{RMSE}{Y_{obs,max} - Y_{obs,min}} \tag{A.5}$$

where, $Y_{obs,max}$ is the maximum value of observed data and $Y_{obs,min}$ is the minimum value of observed data.

5. References

[1] Imai C, Armstrong B, Chalabi Z, Mangtani P, Hashizume M. Time series regression model for infectious disease and weather. Environmental Research. 2015;142:319-327.

[2] Xiang J, Hansen A, Liu Q, Liu X, Tong MX, Sun Y, Cameron S, Hanson-Easey S, Han GS, Williams C, Weinstein P. Association between dengue fever incidence and meteorological factors in Guangzhou, China, 2005–2014. Environmental Research. 2017;153:17-26.

[3] He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. International Journal of Infectious Diseases. 2018;74:61-70.

[4] Eisenberg MC, Kujbida G, Tuite AR, Fisman DN, Tien JH. Examining rainfall and cholera dynamics in Haiti using statistical and dynamic modeling approaches. Epidemics. 2013;5(4):197-207.

[5] Du Z, Lawrence WR, Zhang W, Zhang D, Yu S, Hao Y. Interactions between climate factors and air pollution on daily HFMD cases: A time series study in Guangdong, China. Science of the Total Environment. 2019;656:1358-1364.

[6] Christiansen CF, Pedersen L, Sørensen HT, Rothman KJ. Methods to assess seasonal effects in epidemiological studies of infectious diseases—exemplified by application to the occurrence of meningococcal disease. Clinical Microbiology and Infection. 2012;18(10):963-969.

[7] Moirano G, Gasparrini A, Acquaotta F, Fratianni S, Merletti F, Maule M, Richiardi L. West Nile virus infection in Northern Italy: case-crossover study on the short-term effect of climatic parameters. Environmental Research. 2018;167:544-549.

[8] Efendi R, Arbaiy N, Deris MM. A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. Information Sciences. 2018;441:113-132.

[9] BV BP, Dakshayini M. Performance analysis of the regression and time series predictive models using parallel implementation for agricultural data. Procedia Computer Science. 2018;132:198-207.

[10] Mudelsee M. Trend analysis of climate time series: A review of methods. Earth-Science Reviews. 2019;190:310-322.

[11] Croston JD. Forecasting and stock control for intermittent demands. Journal of the Operational Research Society. 1972;23(3):289-303.

[12] Dagum EB. Revisions of time varying seasonal filters. Journal of Forecasting. 1982;1(2):173-187.

[13] Huyot G, Chiu K, Higginson J, Gait N. Analysis of revisions in the seasonal adjustment of data using X-11-

ARIMA model-based filters. International Journal of Forecasting. 1986;2:217-229.

[14] Okouma Mangha V, Ilk D, Blasingame TA, Symmons D, Hosseinpour-zonoozi N. Practical considerations for decline curve analysis in unconventional reservoirs - application of recently developed rate-time relations. In:SPE Hydrocarbon Economics and Evaluation Symposium 2012. Calgary, Alberta, Canada;2012.

[15] Ray BK. Long-range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA model. International Journal of Forecasting. 1993;9(2):255-269.

[16] Taylor SJ. Forecasting the volatility of currency exchange rates. International Journal of Forecasting. 1987;3(1):159-170.

[17] Tsay RS. Regression models with time series errors. Journal of the American Statistical Association. 1984;79:118-124.

[18] Willemain TR, Smart CN, Shockor JH, DeSautels PA. Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. International Journal of Forecasting. 1994;10(4):529-538.

[19] Asai P, Panja P, McLennan J, Moore J. Performance evaluation of enhanced geothermal system (EGS): Surrogate models, sensitivity study and ranking key parameters. Renewable Energy. 2018;122:184-195.

[20] Wu B, Zhang X, Jeffrey RG. A model for downhole fluid and rock temperature prediction during circulation. Geothermics. 2014;50:202-212.

[21] Hadgu T, Kalinina E, Lowry TS. Modeling of heat extraction from variably fractured porous media in enhanced geothermal systems. Geothermics. 2016;61:75-85.

[22] Mudunuru MK, Karra S, Harp DR, Guthrie GD, Viswanathan HS. Regression-based reduced-order models to predict transient thermal output for enhanced geothermal systems. Geothermics. 2017;70:192-205.

[23] Spiess AN, Neumeyer N. An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC pharmacology. 2010;10(1):6.

[24] Pesaran MH, Smith RP. Signs of impact effects in time series regression models. Economics Letters. 2014;122(2):150-153.

[25] Nottingham QJ, Cook DF. Local linear regression for estimating time series data. Computational Statistics & Data Analysis. 2001;37(2):209-217.

[26] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. International Journal of Forecasting. 2006;22(3):443-473.

[27] Muth JF. Optimal properties of exponentially weighted forecasts. Journal of the American Statistical Association. 1960;55:299-306.

[28] Gardner Jr ES. Exponential smoothing: The state of the art. Journal of Forecasting. 1985;4(1):1-28.

[29] Snyder RD. Recursive estimation of dynamic linear models. Journal of the Royal Statistical Society. Series B (Methodological). 1985;47(2):272-276.

[30] Yule GU. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character. 1927;226:267-298.

[31] Box GE, Jenkins GM. Time series analysis: forecasting and control. Holden-Day; 1970.

[32] Kalman RE. A new approach to linear filtering and prediction problems.Journal of Basic Engineering. 1960; 82:35-45.

[33] Schweppe F. Evaluation of likelihood functions for Gaussian signals. IEEE Transactions on Information Theory. 1965;11(1):61-70.

[34] Shumway RH, Stoffer DS. An approach to time series smoothing and forecasting using the EM algorithm. Journal of Time Series Analysis. 1982;3(4):253-264.

[35] Wiener N. Nonlinear problems in random theory. Cambridge: Technology Press of Massachusetts Institute of Technology; 1958.

[36] Volterra V. Theory of functionals and of integral and integro-differential equations. Blackie & Son Limited; 1930.

[37] Ray BK. Modeling long-memory processes for optimal long-range prediction. Journal of Time Series Analysis. 1993;14(5):511-525.

[38] Bollerslev T, Engle RF, Nelson DB. Chapter 49 ARCH models. Handbook of Econometrics. 1994;4: 2959-3038.

[39] Engle RF. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society. 1982;50:987-1007.

[40] Arps JJ. Analysis of decline curves, SPE-945228-G. Transactions of the AIME. 1945;160.