Rainfall Forecasting using Support Vector Regression Machines

Lemuel Clark Velasco[1]*, Johanne Miguel Aca-ac[2], Jeb Joseph Cajes[3], Nove Joshua Lactuan[4], Suwannit Chareen Chit[5]
Premier Research Institute of Science and Mathematics[1]
MSU-Iligan Institute of Technology, Iligan City, The Philippines[1, 2, 3, 4]
Universiti Utara Malaysia, UUM Sintok, Kedah, Malaysia[5]

***Abstract*—Heavy rainfall as a consequence of climate change have immensely impacted the ecology, the economy, and the lives of many. With the variety of available predictive tools, it is imperative that performance analysis of rainfall forecasting models is properly conducted as a measure for disaster preparedness and mitigation. Support Vector Regression Machine (SVRM) was utilized in predicting the rainfall of a city in a tropical country using a 4-year and 17-month rainfall dataset captured from an automated rain gauge (ARG) in Southern Philippines, involving parameter cost and gamma identification to determine the relationship between past and present values, determining optimal cost and gamma parameters to improve prediction accuracy, and forecasting model evaluation. The SVRM model that utilized Radial Basis Function (RBF) kernel function having the parameters of c=100; g=1; e=0.1; p=0.001 and the lag variable which used 12-hour report with lags up to 672-timesteps (i-672) demonstrated a Mean Square Error (MSE) of 3.461315. With close to accurate forecast between the predicted values and the actual rainfall values, the results of this study showed that SVRM has the potential to be a viable rainfall forecasting model given the proper data preparation, model kernel function selection, model parameter value selection and lag variable selection.**

Keywords*—Support vector regression machines; support vector machines; rainfall forecasting

I. Introduction

Climate change is a widespread and growing threat to biodiversity and ecosystems globally. One of the disasters caused by climate change is heavy rainfall and its frequency is noticeable among tropical countries resulting to catastrophic disasters such as landslides and flood which led to loss of lives, property, and livestock. Rainfall forecasting has received immense attention in recent years due to heightened emphasis on minimizing life and property losses through proper conduct of mitigation and preparedness in disaster risk reduction [1]–[3]. Rainfall forecasting models need to be evaluated and optimized for efficient performance in order for these predictive models to be utilized as disaster risk management tools that can serve as decision-making tools to alert individuals on incoming natural disasters through advance notice for the tactical planning of activities and approach [2]. Support Vector Machines (SVM) with a specific forecasting variant known as Support Vector Regression Machines (SVRM) is an emerging high performing machine learning algorithm used for natural phenomena such as rainfall forecasting [4], [5]. SVRM finds a hyperplane in a n-dimensional space with n-number of features that specifically classifies the data points into classes, applying structural risk minimization elementary principles to obtain quality generalization on a $_{\text{finite}}$ number of learning patterns [3], [5]–[8] . Since SVRM first plots the inputs into a high-dimensional space and looks for a parting hyperplane that maximizes the margin between the classes which then uses kernels to find the optimal hyperplane, various SVRM models can be developed, evaluated and optimized in order to be efficiently used as forecasting models to be effectively used disaster risk reduction.

Iligan City in Southern Philippines, has recently been in the pathway of typhoons due to climate change. Local authorities led by the Iligan City Disaster Risk Reduction Management Council (CDRRMC) need to develop decision support tools in disaster mitigation, preparedness, response and recovery. Despite availability of historical rainfall data, there is a lack of a forecasting model to determine rainfall which is essential in the resource and mitigation planning in times of disasters. Additionally, proper rainfall data preprocessing as well as optimal parameter configuration of forecasting models are needed to utilize rainfall forecasting tools that yields reliable predictive results [2], [7]. This study attempted to conduct performance analysis of SVRM models by conducting rainfall data preparation, kernel selection, SVRM parameter selection, lag variable selection, and implementation of SVRM models. Through model verification in terms of error computation, an assessment on the predictive performance of these SVRM models was conducted to determine the reliability of the forecasting model. By conducting performance analysis, this study hopes to contribute in the on-going efforts to evaluate and optimize efficient performance of rainfall forecasting models that can be used as decision-making tools to save both lives and property.

II. Methodology

A. Rainfall Data Preparation

The rainfall data preparation phase of this study involves three activities namely data selection, data correction and data representation. Data selection is the process of determining the appropriate dataset input variable along with its corresponding time-range as input layer data [3], [9]. The researchers considered rain, rain rate, air temperature, humidity, air pressure, water level and solar radiation as candidate meteorological data that can be possible datasets to be selected [1], [2], [9]. Input data for this study were obtained from the

*Corresponding Author.

Philippine Department of Science and Technology - Advance Science and Technology Institute (DOST-ASTI) which were collected rainfall data from Automated Rain Gauges (ARG) having 15-minute intervals with a unit of measurement in millimeters. The acquired dataset may have incomplete, inconsistent, and noisy values which can be due to incomplete data lacking attribute values, lacking certain attributes of interest. Thus, data correction by finding, checking, or eliminating corrupt and inaccurate records from the rainfall dataset was then conducted to discriminate incomplete, incorrect, or irrelevant parts of the dataset. Manual visual inspection of the raw rainfall data from the obtained spreadsheet file was initially conducted to determine the extent of the data cleaning as well as data recording anomalies such as missing rainfall data brought about by ARG limitations. Further examination of the dataset was performed that examined aggregate data, noisy data containing errors or outliers, and inconsistent data containing discrepancies in codes or names [3], [10].

As shown in Fig. 1, the acquired rainfall data is device-dependent that is why data representation which involved representation processing of key variables and attributes was then conducted. The raw dataset also contains the location of the ARG (LOCATION), latitude of the ARG (LATITUDE), longitude of the ARG (LONGITUDE), elevation of the ARG (ELEVATION), date of installation of the ARG (DATE INSTALLED) and date of reading and amount of rainfall (DATE/TIME READ; RAINFALL AMOUNT). The rainfall data captured in 15-minute observations in terms of millimeters along with its date of reading was considered in this study as the training, testing and validation data sets following the data partitioning process.

```
@relation RogongonOriginal
@attribute Date date "MM/dd/yyyy HH:mm"
@attribute RAINFALL_ROGONGON numeric
@data'03/01/2013 0:00',0
'03/01/2013 0:15',0
'03/01/2013 0:30',0
'03/01/2013 0:45',0
'03/01/2013 1:00',0
```

Fig. 1. Sample Raw Data from the ARG.

B. SVRM Model Design

The design of an SVRM forecasting models depends on a set of specific selection processes which includes kernel selection, parameter selection, and architecture selection [2], [3], [6], [10]–[12]. As shown in Fig. 2, SVRM utilizes kernel functions to transform and map training data from an input space into a high dimensional feature space in which it searches for an optimal classification hyperplane that separates the data into different categories [10], [12]–[14]. Configured by the researchers, the kernel function is the component of the SVRM that plays a central part in the assimilation of data and transforming such data for pattern discovery and general types of relations such as classifications, rankings, cluster, and regressions [13], [14]. Kernel selection as conducted by the researchers involves the testing of different kernel functions to determine the optimal parameters needed to build the SVRM model. The data set was tested with Linear Function Kernel–which excels in linearly separable data, and Radial Basis Function (RBF) Kernel–which is excellent with nonlinear data sets. The study followed proposed the methodology to come up with the best fitted kernel function for an SVRM rainfall forecasting model [15]. The process involved utilizing the grid search method to produce an optimal parameter which was used as a reference parameter for the kernel selection process in an integrated development environment (IDE) using the entire training set and evaluated on the testing set. The kernel functions were then observed for their forecasting accuracy and behavior with the model having the lowest Mean Squared Error (MSE) adopted and used in the parameter selection phase.

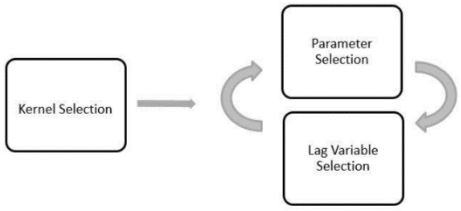

Fig. 2. Model Design Flow.

As supported by studies which used the MSE error metric in assessing the accuracy of the SVRM models, the selection for which parameters was included in model revolved around the selected kernel [16], [17]. The researchers then conducted parameter selection where optimal values of key parameters are selected for forecasting unknown data [4], [11], [12], [16], [17]. The process involves utilizing the Grid Search method to find an optimal combination value of the key parameters Cost (*C*), Gamma (*g*), and Epsilon (*e*). Table I shows that the key parameter (*C*) defines the penalty for errors, the parameter (*g*) influences the hyper-line flexibility, while the parameter (*e*) defines the upper and lower bound of the fractions of the support vectors relative to the total number of training examples [3], [6], [10]–[12], [17].

Utilizing the selected kernel from the kernel selection process, testing of different parameter values were determined with the use of Grid Search and an Exhaustive Parameter Search methodology using the data training and testing sets. This is to validate the reliability of Grid Search in searching for optimal parameter values and to also find the best fitted parameters for the final model [3], [15]. The study also utilized a feature space which was used in both Grid Search and Exhaustive Parameter Search methodologies. The parameter values were evaluated and observed for their forecasting accuracy. Parameter values with the lowest MSE were selected as the optimal parameter values for the kernel function and final forecasting model. The kernel adopted along with the selected parameter values were used to select the lag variable that were used in the final model.

TABLE I. PARAMETERS AND ITS FUNCTION

PARAMETERS	FUNCTION
Cost Penalty (C)	Defines the penalty for errors
Gamma (g)	Influences the hyper-line flexibility
Insensitive Loss Function (e)	Width of ε-insensitive zone/tube
Epsilon (p)	The set of epsilon function in epsilon SVR

Lag variable selection, where optimal value for the lagged variables were selected, was used by the researchers to determine the relationship between past and current values of the series which can be captured by a propositional learning algorithm like the SVRM algorithm. Fundamentally, the value of the lagged variables created regulates the size of the time window [6], [10]. This step involves evaluating 2 lag variable values wherein each lag value utilized 4 years, 7 months, and 30 days' worth of rainfall data in periods of 15 minutes. The first lag variable value (Lag Variable I) had a per-12-hour report with lags up to 672-time steps (i-672) in the past, wherein *i* represents the current date of the model. The configuration of the first lag variable value was considered since the heuristic standard for weather forecasting is on a weekly basis. The second lag variable value (Lag Variable II) had a per-12-hour report with lags up to 2976-time steps (i-2976) in the past. The configuration of the second lag variable value was considered since rainy season in countries like the Philippines starts in June and last till November, wherein the months of September and October are often the typhoon season in the entire archipelago. The forecasted values were validated using the remaining month of the data. Using the selected kernel function, kernel parameter values, and the data set, architectures determined were used through a determined IDE. The lag variables were observed for its forecasting accuracy and behavior with the lag variable having the lowest MSE was selected for the final model.

C. SVRM Model Evaluation

The researchers integrated and constructed an SVRM model utilizing the LIBSVM library for SVRM processes to improve efficient and effective management of the rainfall forecasting process. It is necessary that the computing environment settings befit for developing SVRM model and the needed libraries and IDE should be ready before the development start. Computer running on Mac OS, Windows, or Ubuntu is necessary, with an IDE, preferably the Waikato Environment for Knowledge Analysis (WEKA). Few WEKA packages or extension were installed like grid search that handled optimization of the parameters and enhance the specific values for the parameters. The SVRM programme stimulates the training, validating, and forecasting of the data after creating the data and environment in which the system was constructed. The specification of parameters, kernel, lag variable, and training of the prepared dataset was then referred to as training. Furthermore, the trained model was validated by using an error metric until an output was finally generated.

Using the testing set, SVRM model evaluation examined each model in the training process for its accuracy in forecasting 12 hour-ahead rainfall values in 672-time or 2976-time lagged steps in the past. The predicted result was compared to the testing set's actual rainfall data. For the rainfall forecasting system, the model with the lowest MSE was then chosen as the final model. The computation of the measure of error is an important aspect of evaluating a forecasting model's prediction accuracy. The forecasting error, which is the difference between the anticipated and actual rainfall levels, is a measure of a forecasting model's accuracy [17]. The MSE shown in Equation (1) was used in this study since the actual values of the data are in the denominator of the equation and will produce undefined or infinite results when the actual demand is zero.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i^v)^2 \quad (1)$$

where Y_i are the observed values of the variable being predicted and $\hat{Y}_i$ are the predicted values.

The MSE, the model's consistency will be indicated by a minimal error. The rainfall data from the testing set was then used to test the selected SVRM models. The data was fed into the chosen SVRM models, which generated anticipated results for the next 12 hours. In a 15-minute cycle, 80% of the rainfall data was loaded into the selected models to anticipate rainfall for the next 12 hours. The predicted values were then compared to the testing set's actual rainfall value. The anticipated rainfall values were evaluated using MSE after the evaluation phase was completed. Error assessment was then conducted where the validation results were converted and compared to the actual results to see if they are accurate in terms of the actual rainfall data [3], [11] . In this process, the validation set was utilized to generate validation findings throughout the month of October in 2017. The accuracy of the forecasting was then tested by comparing these findings. Following the selection of the best performing model, the forecasted values were graphed into a line graph and compared to the actual rainfall levels. For charting, the values of each iteration of each model were aligned with the real values of the same week. Visual inspection was then carried out by watching it on a weekly basis. Every day in October, for example, the next seven days were gathered and graphed. As a result, the full validation set was observed for a total of 31 days. The t-score shows the differences of the two groups, the larger values specify the difference between the two samples while the smaller values specify the similarity of the two samples. The t-score *t*, gave the means of the first and second samples, X_1 and X_2, with n being the sizes of both X_1 and X_2 with S_p the pooled standard deviation is shown in Equation (2).

$$t = \frac{x_1 - x_2}{s_p\sqrt{\frac{2}{n}}} \quad (2)$$

In this study, X_1 and X_2 are the means of the forecasted and actual values for rainfall in each week. The pooled standard deviation S_p provided the standard deviations in samples S_1^2 and S_2^2, with n_1 and n_2 representing the sample size of the first and second group as provided by Equation (3).

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}} \quad (3)$$

III. Results and Discussion

A. Rainfall Data Preparation Results

In the preparation of the rainfall data for the SVRM model design, the rainfall data used in the study was selected, cleaned, represented and partitioned. Rainfall Data from 2013 to 2017 was exported as a .csv file from the ARG installed in Rogongon, Iligan City (Latitude: 8.232697, Longitude: 124.419372) on February 12, 2013. The attributes referring to location (LOCATION), latitude (LATITUDE), longitude (LONGITUDE), elevation (ELEVATION), and date of

installation (DATE INSTALLED) were removed since it plays insignificant contribution to the learning process. A delimiter on semicolon was then used to separate the attribute of a single data unit for the date, time of reading and the rainfall amount (DATE/TIME READ; RAINFALL AMOUNT). Shown in Table II is the format for the 163, 777 rows of data having its respective date, time, and rainfall. The dataset was partitioned into 80% training set and 20% testing set and was then converted into .arff to be suited for WEKA.

TABLE II. FORMAT OF THE DATASET

DATE / TIME	RAINFALL
NN/NN/NNNN 00:15	NN
NN/NN/NNNN 00:30	NN
NN/NN/NNNN 00:45	NN
NN/NN/NNNN 01:00	NN

B. SVRM Model Design Results

Kernel functions plays a central part in the assimilation of data and transforming such data for pattern discovery. Thus, which kernel would be most suitable depends upon the data that the kernel would use [13], [18]. The study evaluated the Linear Function kernel, which is an excellent kernel function for linearly separable data sets, and the RBF kernel, which excels with nonlinear data sets [3], [6], [13]. To determine if a Linear or RBF kernel function will be used in the final model, the researchers evaluated the MSE value the function kernels were ableyield. Table III shows the results of the kernel selection process which the study utilized the parameter values produced by Grid Search as reference parameter values.

TABLE III. GRID SEARCH RESULT

COST	GAMMA	ACCURACY
100	10	91%

As shown in Table IV, the Cost value of 100 and the Gamma value of 10 were obtained through the Grid Search procedure. The accuracy was calculated using an 80/20 split, with 80% of the training set and 20% of the testing set yielding a 91% accuracy on the testing set. To calculate the MSE of the Linear and RBF kernels, the Cost and Gamma values obtained during the Grid Search procedure were then used as reference parameter values.

TABLE IV. COMPARISON BETWEEN LINEAR AND RBF KERNELS

LINEAR PARAMETER	LINEAR MSE VALUE	RBF PARAMETERS	RBF MSE VALUE
Cost (C) = 100	MSE > 100	Cost (C) = 100 Gamma (g) = 10 Loss Function (e) = 0.1 Epsilom (p) 0.001	3.6377

Kernel parameters were employed in nonlinear feature mapping to govern the trade-off between margin maximization and error minimization. The hyper parameters regulate the model's training process and have a significant impact on the SVRM forecasting model's development and test performance [12], [18]. The hyper parameters control the training process of the model and have an extensive effect in the development and resulting test performance of the SVRM forecasting model. The results shows that RBF kernel yielded an MSE of 3.6377 while the linear kernel yielded an extensive MSE value of greater than 100. This could only mean that the data set was not linearly separable but is nonlinear in nature. In the behavior of the data used in this study, RBF was found to be more accurate. As such, RBF kernel function was the choice for the SVRM modelling.

This study compared Grid Search and Exhaustive Search methodologies to determine the best fit parameter combination and to validate Grid Search methodology's reliability and compatibility with the LibSVM library and WEKA forecasting software. After conducting Grid Search and Exhaustive Search testing, the parameter combination values MSE's were recorded, and the results are presented in Table V and VI. The performance of an SVRM forecasting model relies on three key parameters. The key parameter (*C*) is a parameter that allows the trade-off between training error and model complexity, parameter (*C*) defines the penalty for errors. If the value of parameter (*C*) is too big there would be the likelihood of overfitting a model. Whilst having a smaller (*C*) parameter value may result to the underfitting of a model and increase the number of training errors. The parameter (*g*) influences the hyper-line flexibility while parameter (*e*) controls the width of the ε-insensitive zone, used to fit the training data. If the value of parameter (*e*) is big, this will result in having fewer support vectors selected, and will result in a flatter or less complex regression estimates. Results of the Grid Search produced a Cost of 100 and a Gamma of 10 and yielded an MSE of 3.637. While Exhaustive Search produced a Cost of 100 and a Gamma of 1 and yielded an MSE of 3.6377 as observed in Table IV. Although Exhaustive Search yielded the lower MSE among the two methodology's the difference between the two is only 0.0004, a small difference. From the tested models shown, it can be noted that both parameters set values can be applied in the final model. The researchers opted for the parameter results yielded by the exhaustive search with c=100; g=1; e=0.1; p=0.001 for the final model.

TABLE V. CONSTANT PARAMETER VALUES

INSENSITIVE LOSS (e)	EPSILON (p)
0.1	0.001

TABLE VI. GRID SEARCH RESULT

COST	GAMMA	MSE
100	10	3.637

The researchers provided a feature range of exponents for the Grid Search. For Cost (C) = {0.0001, 0.001, 0.01, 0.1, 1, 10, 100} and for Gamma (g) = {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000}. The search was tested for all 56 combinations of Cost and Gamma. Grid Search is tested for each combination, evaluating for each of the combinations MSE. The Grid Search process ended with the parameter combination of Cost (C) = 100 and Gamma (g) = 10, with an MSE of 3.637, the lowest MSE determined by the Grid Search process out of the 56 possible combinations. The researchers also tested tuning the parameter using a more practical approach to the process. Like Grid Search, the researchers tested all 56 combinations for

Cost and Gamma (C={10-4, 10-3…101,102}; g={10-4, 10-3…102,103}) manually using MSE to find the combination with the lowest MSE. The search ended with parameter combination of Cost (C) = 100 and Gamma (g) = 1, with an MSE of 3.6366. The procedure entails analyzing two lag variable values, each of which will use rainfall data from the previous four years, seven months, and thirty days in 15-minute intervals. The first lag variable value (Lag Variable I) produced a per-12-hour report with lags up to 672-time steps (i-672) in the past, with I representing the model's current date. The second lag variable value (Lag Variable II) provided a per-12-hour report with lags as far back as 2976-time steps (i-2976). The testing set was used to determine the lag values. The MSE result for each lag variable value in each 15-minute cycle is shown in Table VII for each 12-hour report. The accuracy of Lag Variable I is better. Although Lag Variable I had the lower MSE score, it is worth noting that the difference between it and Lag Variable II is only about 0.00247.

TABLE VII. 12 HOUR-AHEAD MSE VALUES

DATE / TIME	LAG VARIABLE I	LAG VARIABLE II
10/31/2017 0:00	3.6387	3.6388
10/31/2017 0:15	3.6387	3.6389
10/31/2017 0:30	3.6387	3.639
10/31/2017 0:45	3.639	3.6391
---	---	---
---	---	---
10/31/2017 23:15	3.6436	3.6436
10/31/2017 23:30	3.6437	3.6437
10/31/2017 23:45	3.6438	3.6439
MSE	**3.64129**	**3.64154**

C. SVRM Model Evaluation

The optimal lagged variable was chosen after thorough test and comparison. The results show that Lag Variable I with the value of i-672 yields slightly better results than Lag Variable II (i-2976) making it a viable option for a model. The lowest MSE is achieved after tuning the cost value to 100 and gamma value to 1 which is optimal based on the training set used. The test shows a 0.0247 difference between Grid Search using RBF kernel with 3.6377 MSE value compared to 3.6366 MSE value of manual tuning. The training set produced an average of 3.641315 MSE value with minimum and maximum value 3.6387 and 3.6439 respectively using Lag Variable I. The values produced were relatively high for MSE due to the nature of the training set. As shown in Table VIII, the data yielded a t-value of 3.95426E-74. The p-value p = 0.05 by default which means the forecasted data is acceptable if the margin of error is less than 5%. The researchers used a total of 58 data points by using 29 data points from forecasted and actual data sets. The degrees of freedom df = 56 (degrees of freedom is number of datapoints minus 2) with its critical value 1.673 is used in assessing the H_0 where is H_0 is not rejected since the t-value is less than the critical value. This shows that the predicted values have no statistically significant difference from the actual values. The low t-value also means that the difference between the actual and forecasted values is extremely small and the error becomes insignificant.

TABLE VIII. SNAPSHOT OF THE STUDENT'S T-TEST

DATE / TIME	ACTUAL	PREDICTED
10/01/2017 0:00	2.434331	2.3236
10/01/2017 0:15	2.434833	2.3235
10/01/2017 0:30	2.435335	2.3235
10/01/2017 0:45	2.435837	2.3234

Visual representation of the rainfall data exhibits the minimal differences between the actual and forecasted values from the validation set with the lines of the graphs overlapping except for the time intervals with the values equal to 0. The pattern being generated in Fig. 3 shows that the average predicted rainfall values were close to the actual rainfall values. This indicates that the model was able to create an accurate prediction result for the average rainfall value for the validation set [3], [9], [10], [15]. However, having an accurate prediction result in the first and middle half of the data set does not mean that the prediction accuracy will not drop. It is worth noting the fact that regardless of the magnitude of the error, an error will still propagate further errors which will eventually drop the prediction accuracy further down the timeline, especially when the number of units of the predicted values are overstretched. Shown in Fig. 4, the pattern generated shows that a significant error ensued in the beginning of the prediction result. The error continued along the timeline further dropping the accuracy of the result. The model was not able to accurately predict the maximum rainfall value for the data set. This indicate that the model was not able to predict certain change in the data value which in this case are the sudden increase in the value of rainfall due to sudden heavy rain down pour.

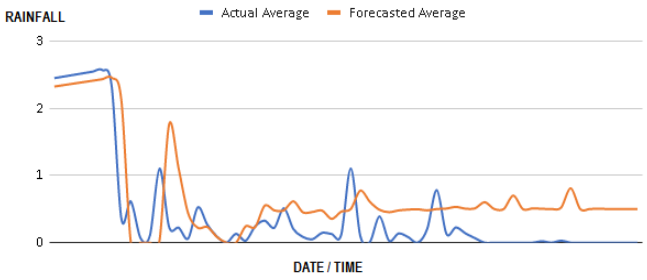

Fig. 3. Average Actual and Forecasted Rainfall.

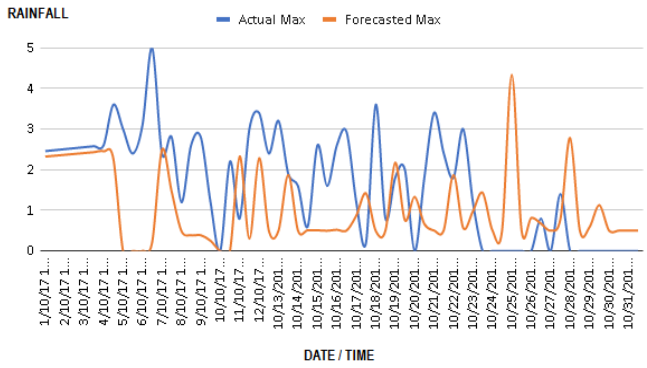

Fig. 4. Maximum Actual and Forecasted Rainfall.

IV. Conclusion and Recommendations

This study attempted to implement a rainfall forecasting strategy using SVRM by performing data preparation, SVRM model design, model implementation and testing the forecasted results for performance evaluation and model validation. On the data preparation process, data correction and representation of the dataset greatly affects the outcome of the data being predicted. Manual vision inspections were conducted and were able to remove irrelevant dates with missing values which results into a number of 163,777 rows. In SVRM model design, it was found out that in order to produce a good forecasting outcome, the right values must be identified for parameters cost and gamma, along with a kernel function that will fit the data set along with a lag variable value that can optimally determine the relationship between the past and current values of the data set. The rainfall dataset was tested with both linear and RBF kernel functions. The data set was first tested with RBF kernel function with temporary base parameter values for cost and gamma identified with the use of Grid Search, the temporary base parameter values were Cost (C) = 100 and Gamma (g) = 10 with an accuracy of 91% resulting to a MSE value of 3.6377. The data set was then tested with the linear kernel function using the same base parameter values and resulted in an MSE value greater than 100. The researchers concluded that the data set was non-linear in nature and is not linearly separable. Thus, the model utilized RBF as its selected kernel function. The second phase of the selection involved selecting the best parameter values for cost and gamma. The study utilized two tuning techniques for the selection process; the following were the Grid Search and Exhaustive Search techniques. Tuning was first tested using Grid Search, the search produced values of Cost (C) = 100 and Gamma (g) = 10 with an MSE value of 3.637. The second tuning was then tested using Exhaustive Search with a feature space range for Cost (C) = {0.0001, 0.001, 0.01, 0.1, 1, 10, 100} and Gamma (g) = {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000}. Exhaustive search produced the lower MSE value of 3.6366 with values for Cost (C) = 100 and Gamma (g) = 1. The results yielded in only an MSE value difference of 0.0004, a very small difference. It can be concluded that both tuning techniques can be utilized for tuning parameters in the creation of the SVRM model design. However, the researchers opted for the parameter set values yielded by Exhaustive Search parameter tuning in the final model. The last phase of the selection process involved the selection of a lag variable value. The lag variable determines the past and current relationship of a data series in a particular timeframe. The study tested two variable values; Lag Variable I and Lag Variable II. Lag Variable I have a per-12-hour report with lags up to 672-timesteps (i-672) in the past, while Lag variable II has a per-12-hour report with lags up to 2976-time steps (i-2976) in the past. Lag variable I was first tested and produced an MSE value of 3.641315, while Lag variable II produced an MSE value of 3.651315 in the second test. Though Lag Variable I yielded the lower MSE score, it is also worth noticing that it only differs by a small variance of approximately 0.00247 when compared with Lag Variable II. This is an indication that Lag Variable II is also a promising Lag Variable value for the SVRM forecasting model. The final model utilized Lag variable I for the final model.

It is recommended that a different rainfall dataset from a non-tropical country be used to validate the SVRM models presented in this study. Having datasets with a vast difference of rainfall values is expected affect the performance of the model which in turn affects the accuracy, behavior and performance of the SVRM. Tropical climate like that of the Philippines having only wet and dry seasons anytime within the year records a different rainfall behavior from geographies having four seasons. The researchers would also like to recommend for further studies on the aspect of kernel, lag variable and architecture selection. Further studies on these processes will help optimize the performance of the SVRM in rainfall forecasting. Aside from WEKA, other SVRM development frameworks could also be used to expand model performance analysis conducted in this research. One or more SVRM development frameworks can be compared with the model results of presented in this study as well as conducting a contrast if other development frameworks have better or the same performance with that of WEKA. Overall, the results of this study showed that SVRM has the potential to be a viable rainfall forecasting model given the proper data preparation, model kernel function selection, model parameter value selection and lag variable selection.

Acknowledgment

The authors would like to thank the support of the Mindanao State University-Iligan Institute of Technology (MSU-IIT) Office of the Vice Chancellor for Research and Extension for their assistance in this study. This work is supported by MSU-IIT as an internally funded research under the Premier Research Institute of Science and Mathematics (PRISM)- Applied Mathematics and Statistics (AMS) Research Group. The authors would also like to thank the Philippines Department of Science and Technology - Advance Science and Technology Institute (DOST-ASTI) for the data used in this study.

References

[1] J. M. West et al., “U.S. Natural Resources and Climate Change: Concepts and Approaches for Management Adaptation,” Environmental Management, vol. 44, no. 6, p. 1001, 2009, doi: 10.1007/s00267-009-9345-1.

[2] R. Muhammad and J. Mahmmud, “Rainfall Event Analysis for Urban Flooding Study Using Radar Rainfall Data,” 2015. [Online]. Available: www.jzs.uos.edu.krd.

[3] N. Hasan, N. C. Nath, and R. I. Rasel, “A support vector regression model for forecasting rainfall,” in 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT), 2015, pp. 554–559. doi: 10.1109/EICT.2015.7392014.

[4] M. Mokhtarzad, F. Eskandari, N. Jamshidi Vanjani, and A. Arabasadi, “Drought forecasting by ANN, ANFIS, and SVM and comparison of the models,” Environmental Earth Sciences, vol. 76, no. 21, p. 729, 2017, doi: 10.1007/s12665-017-7064-0.

[5] A. Pozdnoukhov, G. Matasci, M. Kanevski, and R. S. Purves, “Spatio-temporal avalanche forecasting with Support Vector Machines,” Natural Hazards and Earth System Sciences, vol. 11, no. 2, pp. 367–382, 2011, doi: 10.5194/nhess-11-367-2011.

[6] D. Boswell, “Introduction to Support Vector Machines,” 2002.

[7] Y. Lin, Y. Lee, and G. Wahba, “Support Vector Machines for Classification in Nonstandard Situations,” Machine Learning, vol. 46, no. 1, pp. 191–202, 2002, doi: 10.1023/A:1012406528296.

[8] C.-H. Wu, J.-M. Ho, and D. T. Lee, “Travel-time prediction with support vector regression,” IEEE Transactions on Intelligent

Transportation Systems, vol. 5, no. 4, pp. 276–281, 2004, doi: 10.1109/TITS.2004.837813.

[9] A. El-Shafie, M. Mukhlisin, A. A. Najah, and M. R. Taha, "Performance of artificial neural network and regression techniques for rainfall-runoff prediction," International Journal of Physical Sciences, vol. 6, no. 8, pp. 1997–2003, Apr. 2011, doi: 10.5897/IJPS11.314.

[10] J. Du, Y. Liu, Y. Yu, and W. Yan, "A prediction of precipitation data based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) algorithms," Algorithms, vol. 10, no. 2, Jun. 2017, doi: 10.3390/a10020057.

[11] E. G. Ortiz-García, S. Salcedo-Sanz, and C. Casanova-Mateo, "Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data," Atmospheric Research, vol. 139, pp. 128–136, 2014, doi: https://doi.org/10.1016/j.atmosres.2014.01.012.

[12] J. Zhang, X. Qiu, X. Li, Z. Huang, M. Wu, and Y. Dong, "Support Vector Machine Weather Prediction Technology Based on the Improved Quantum Optimization Algorithm," Computational Intelligence and Neuroscience, vol. 2021, p. 6653659, 2021, doi: 10.1155/2021/6653659.

[13] J. July, D. Ben, A. Mezghani, S. Zribi Boujelbene, and N. Ellouze, "Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification."

[14] H. Fizazi Izabatene, W. Benhabib, and S. Ghardaoui, "Contribution of Kernels on the SVM Performance," Journal of Applied Sciences, vol. 10, no. 10, pp. 831–836, 2010.

[15] G. Adhani, A. Buono, and A. Faqih, "Optimization of Support Vector Regression using Genetic Algorithm and Particle Swarm Optimization for Rainfall Prediction in Dry Season," TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 12, no. 11, Nov. 2014, doi: 10.11591/telkomnika.v12i11.6518.

[16] M. v Shcherbakov, A. Brebels, A. Tyukov, T. Janovsky, and V. Anatol, "A Survey of Forecast Error Measures," 2013.

[17] H. Wang and D. Xu, "Parameter Selection Method for Support Vector Regression Based on Adaptive Fusion of the Mixed Kernel Function," Journal of Control Science and Engineering, vol. 2017, p. 3614790, 2017, doi: 10.1155/2017/3614790.

[18] A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. A. Benediktsson, "Gradient Optimization for multiple kernel's parameters in support vector machines classification," in IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium, 2008, vol. 4, pp. IV-224-IV–227. doi: 10.1109/IGARSS.2008.4779698.