An Ensemble Method Based Aggregated Model by Analyzing Data of Existing Precipitation Prediction Models*

Ramyaa[1] and Kallol Das[2]

***Abstract*— Most of the existing precipitation prediction models are not predicting well enough. Most of the cases these models are over-predicting. Sometimes the rate of false positive is way high. Again, some models are predicting good in some places and worse in other places, i.e., some of them are good at mountain areas, some of them are good at desert areas etc. The goal of this research is to reduce the error rates of the existing prediction models. There are lots of existing researches going on implementing new models to predict precipitation. But, the false positive rate didn't reduced that much. We propose an ensemble approach to develop a New Aggregated Model to predict precipitation based on the dataset of some existing prediction models.**

***Index Terms*— Machine Learning, Precipitation Prediction, Aggregated Model**

I. INTRODUCTION

Predicting correct amount of precipitation for a particular day is always tough. Existing well established precipitation prediction models are not accurate enough. Even sometimes the error rates are too high that turns the model as a bad prediction model. Lots of research going on to improve the prediction accuracy, i.e., to decrease the error rates.

Basically, most of the research that has been done so far implemented new models to predict precipitation from some real features. Traditional statistical analysis techniques were mostly used previously for precipitation prediction. J. C. Thompson [2] proposed a numerical method to predict precipitation. This prediction model was based on a graphical integration technique by using a number of independent variables. Later machine learning started performing more accurately over traditional statistical analysis.

Wei-Chiang Hong [1] proposed a hybrid model of RNNs and SVMs (named as RSVR) to forecast the precipitation amounts. Chaotic Particle Swarm Optimization (CPSO) algorithm has been used to select the parameters of the SVR model. Selected parameters were used to predict precipitation amount. Theoretically that research was showing significantly small Normalized Mean Square Error rate, but, the predicting forecast for verification data and testing data had right shifted result in the time domain.

Emilcy Hernandez et al. [3] proposed a deep learning architecture for the next day precipitation prediction. In total, forty-seven features, including temperature, humidity, wind direction, pressure, previous rainfalls etc., have been used as input in this research to predict the amount of precipitation for the next day. According to the result of this research, new model is less accurate for days with light rainfall.

Beda Luitel et al. [4] has been evaluated the skill of five Neumaric Weather Prediction (NWP) systems [European Centre for Medium-Range Weather Forecasts (ECMWF), UK Met Office (UKMO), National Centers for Environmental Prediction (NCEP), China Meteorological Administration (CMA), and Canadian Meteorological Center (CMC)]. Five other remote sensing products have been compared in this research. One of the remote sensing products performs better than any other products even for a recent storm, Hurricane Joaquin (2015). NWP models on the other hand was able to identify high amount of rainfall at the shortest lead times, but couldn't perform good at longer lead times.

In summary, many research have been proposed new models from the real weather data by considering a good amount of features e.g., temperature, wind speed, humidity etc. Most of the cases, the error rates for the prediction data of those new models are high. [1] [3] Again, some models are more accurate for days with heavy rainfall, but less accurate for light rainfall. [3] Some models are able to predict for shorter durations, but not for longer durations. [4]

A very few works have been done to improve these models. In other word, many more new models are being proposed instead of trying to improve the existing models. We believe, the more important job is to improve those existing models or to combine multiple existing models to come up with a better models with better performance.

We have rainfall dataset of 39 Prediction Models and a Real Verification dataset for 39 different days in total and for each day we have data for 20 different times. The dataset has been provided by Oklahoma University. **Our approach is to propose a new aggregated model from these existing prediction models' data which can perform better than other existing prediction models.** In other word, our goal is to analyze the error rates of these existing prediction models data and propose better model which has lower error rates than the existing prediction models.

II. PROBLEM STATEMENT

The Research Problem we are dealing with is that all our existing prediction models data have high error rates. Over-predicting rates of these models are quite high in lots of places. Again, different models perform better in different places, worse in other places. Implementing new models from meteorological data by introducing different kind of input features are also not fulfilling the expectation. [3] Here

*This work was supported by Oklahoma University

[1]Ramyaa is with Faculty of Computer Science, New Mexico Tech, 801 Leroy Pl, New Mexico, USA ramyaa.ramyaa at gmail.com

[2]Kallol Das is with the Department of Computer Science, New Mexico Tech, 801 Leroy Pl, New Mexico, USA kalloldash at gmail.com

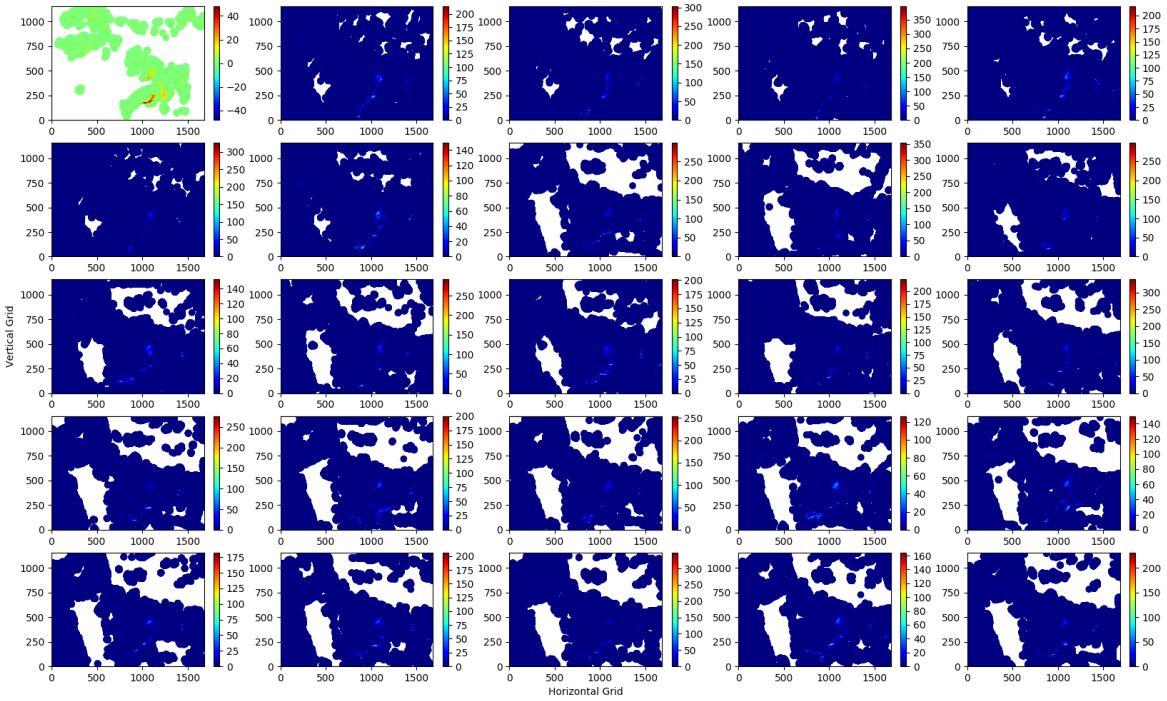

Fig. 1. Visualization of precipitation real verification data and prediction models' data. The first image (image in the top left corner) is the real verification precipitation data and rest of the 24 images are visualizing precipitation prediction data from 24 prediction models.

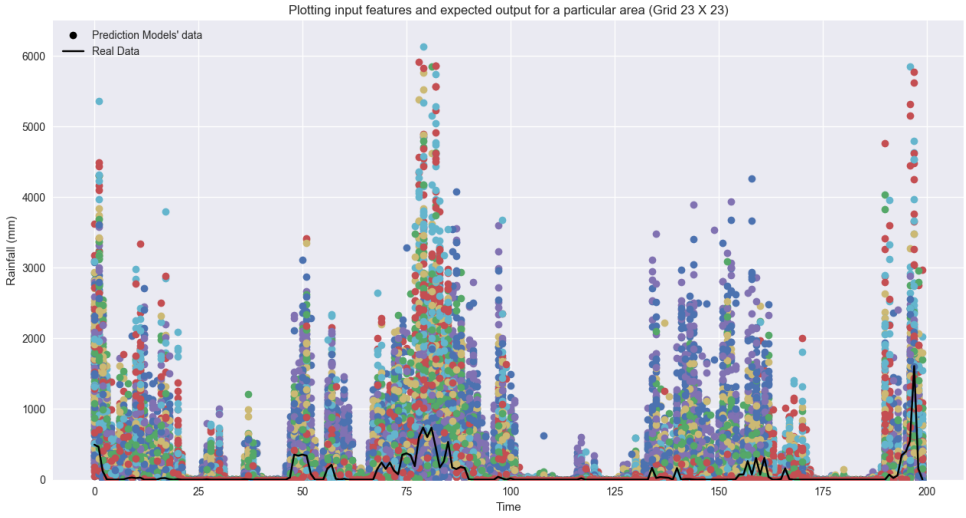

Fig. 2. Plotting input features and real verification data for a particular area which is position 23 X 23 in the grid.

comes our research question- can we implement a new model from these existing models data which has comparatively lower over-prediction and lower false positive rate? Since, different models give wrong prediction in different areas, we assume that an aggregated model would perform better than other existing prediction models.

Fig. 1. illustrates all the prediction models data we have and also the real verification data. All the prediction models data have some areas where real data (image in the top left corner) doesn't have any precipitation in those particular area. In other word, data of all the prediction models contains a huge amount of false positive information. Since, the error rates are way high for all the models data, we can't expect any dramatic change in our new model but, we expect that an aggregated model would be able to minimize a significant amount of error rates and would perform better than all the existing prediction model.

Fig. 2. shows the comparison between the predictions of existing data and the real data for a particular grid point. It illustrates the rate of over-predicting. All the prediction models we have, provides high error rates. Most of the prediction models are forecasting rains for some time while

verification data shows rains tends to zero for those time. Again, the highest rainfall amount for the verification data is around 1500mm while some prediction models are predicting rain around 3000mm which is double than the original data. Again, very few models were forecasting closer to the verification data. Proposing new models from meteorological data is not helping to reduce this kind of high error rates. We believe, our proposing method of developing an aggregated model would help to reduce a large amount of error rates.

III. Methodology

Some prediction models contains data for different days than other prediction models. Again, some prediction models don't contain data for some particular times. The data that has been included in this evaluation has 24 prediction models for 20 days in total and 10 different times for each day.

A. Resolution selection

We have the prediction and verification data for 1155X1683 different geographical coordinates for all over the USA. Since, every coordinates cover around 3 kilometers, the initial goal was to find out a stable resolution for comparatively bigger area which prediction is pretty good. In other word, we wanted to start with a low resolution grid where the error rates are comparatively lower. We choose a 25X25 grid to start with and created a 25x25 summing dataset by adding values of all points for every 25x25 grid. After getting a good result for a 25X25 summing dataset, we tried to make the resolution higher with 15X15 and 5X5 summing dataset.

B. Input Grid

It is assumed that, for a particular area which has high error rate, the neighbor grids reflect closer amount to the target value. So, neighbors grid should help to predict precipitation amount more accurately. That's why a 3X3 grid has been given as a input to predict the middle value of that 3X3 grid. In other word, 9 features from every prediction model's data have been given as input to predict 1 precipitation value for a particular place.

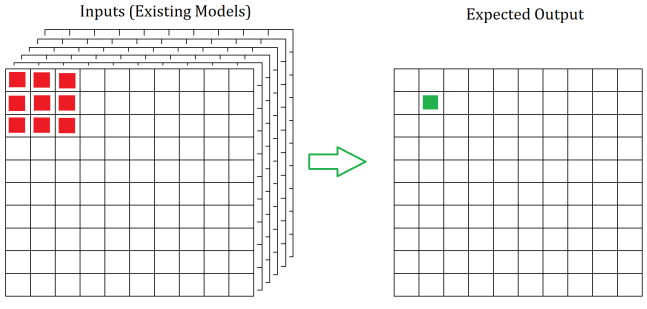

Fig. 3. Visualizing the idea of giving 3X3 grid as input to predict a particular area.

Fig. 3. reflects the idea what we used while fitting data into different machine learning models. 9 red grid illustrate the data from existing prediction models as input and green grid illustrates the expected predicting area. Same idea has been used for all the prediction models' data.

C. Analyzing features with high errors

The Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) for the existing prediction models data have been calculated over all days and times.

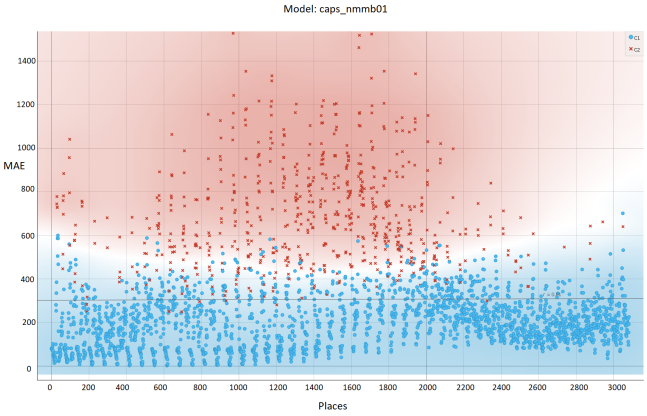

Fig. 4. Shows the clustering of MAE for an existing prediction model data.

Fig. 4. shows the clustering of lower and higher error rates for a single prediction model (model name: caps_nmmb01) data error rates. X-axis is the places, while the y-axis is the rates of error. Basically, the prediction model is showing lower error rates for some places and higher errors in some other places. So, the argument we are trying to make is, this particular model is not good in some places. That's why whenever we are predicting precipitation amount for those places with higher errors, we should not use the data from this model. Which inputs should be provided to predict precipitation for a particular place and which input should not be provided can be identified using a feature selection technique. To find out the best features, a feature selection technique has been used which is discussed in the next section.

D. Univariate feature selection

As we discussed in the previous subsection, each prediction model data has less error rate for some coordinates and high error rates for some other coordinates. Again, we discussed on the second subsection of this methodology section that we input 3X3 grid as input for all prediction models' data, the total input feature became 216 for every sample. But not all 216 features are helping to predict expected value accurately. That's why we decided to use a feature selection technique.

An Univariate Feature Selection technique has been used using Orange Data Mining Toolbox in Python. [6] According to Orange Documentation, univariate linear regression is a univariate feature selection technique for continuous variable that shows the relationship between dependent and independent variable. Orange documentation refers Scikit-Learn Machine Learning Library [5] documentation where

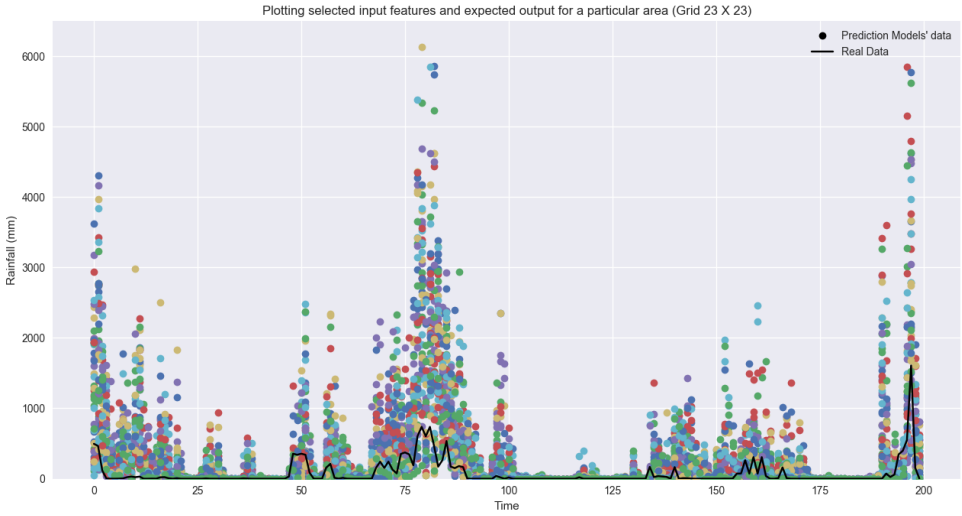

Fig. 5. Visualizing selected 50 features and the target values

they mentioned that the univariate feature selection technique selects feature based on univariate statistical tests.

After using the univariate feature selection technique, most effective 50 features have been kept out of 216 features based on a scoring function. Fig. 5. plots the reduced 50 features input and also the expected output. Before fitting the dataset into machine learning models, a dimensionality reduction technique has been executed.

E. Dimensionality Reduction

Since, the input dataset has only 200 samples, so it is important to reduce the dimension because of the curse of dimensionality. [14] [15] Again, over-fitting [15] should be an important concern. Principal Component Analysis (PCA) has been used to reduce dimension of our input vector. Principal component analysis reduces the dimension of a high dimensional dataset into lower dimension by keeping useful features of data. The goal of using PCA is to summarize the data into limited number of principal components. [7] [8] After using PCA on the existing data, five PCA dimensions have been considered to put as input.

F. Machine Learning Models

Different machine learning models have been used to find out the best prediction. An ensemble technique has been used to take the best models for every grid.K nearest neighbors, Support Vector Machine, Neural Network, Random Forest, Linear Regression have been used to fit data and predict expected output. Best result out of all machine learning models has been taken every time for every single place.

1) K-nearest neighbors: K-nearest neighbors algorithm works for both classification and regression problem. K-nearest neighbors algorithm find the k amount of nearest points which are available in training dataset and predict based on those k neighbors' value. K-nearest neighbors regression basically performs better with less amount of input features. [9] Since we have 5 input features after executing PCA, we expect good result from K-nearest neighbors regression algorithm.

2) Support Vector Machine: Support vector machine is a supervised learning algorithm which works on both classification and regression problems. Basically, the algorithm that works for regression problem is called Support Vector Regression (SVR). SVR works for both linear and non-linear regression problem. The basic idea is to build a hyperplane in a high dimensional space to categorized the input data and predict output. [10]

3) Neural Network: Artificial Neural Network is a machine learning model which consists of an input layer, one or more hidden layers and an output layer. Each neuron of the network works like a information processing unit which is exactly same as a human brain neuron. There are three important types of neural network architectures: Single-Layer Feedforward Networks, Multilayer Feedforward Networks, Recurrent Neural Network. [11] [9]

4) Random Forest: Random forest is an ensemble method which constructs multiple decision tree predictors and predicts output based on voting for the popular class. Random forest perform as good as the bagging and boosting algorithms but runs faster than them. It also calculates the correlation and importance of parameters internally. [9] [12]

5) Linear Regression: Linear regression creates a linear model that represents the relationship between the dependent variable (y) and one or more independent variables (x_1, x_2, ..., x_n). The goal of regression analysis is to define a hypothesis based on the value of dependent and variables and predict y based on the hypothesis. [13]

IV. Experimental Result

All the three summing dataset were able to provide better prediction comparing to the existing models' prediction. For 25x25 summing dataset, 98.44% cases new model's MAE is better than any other existing prediction models. Again, 100.0% cases RMSE of new model is better than

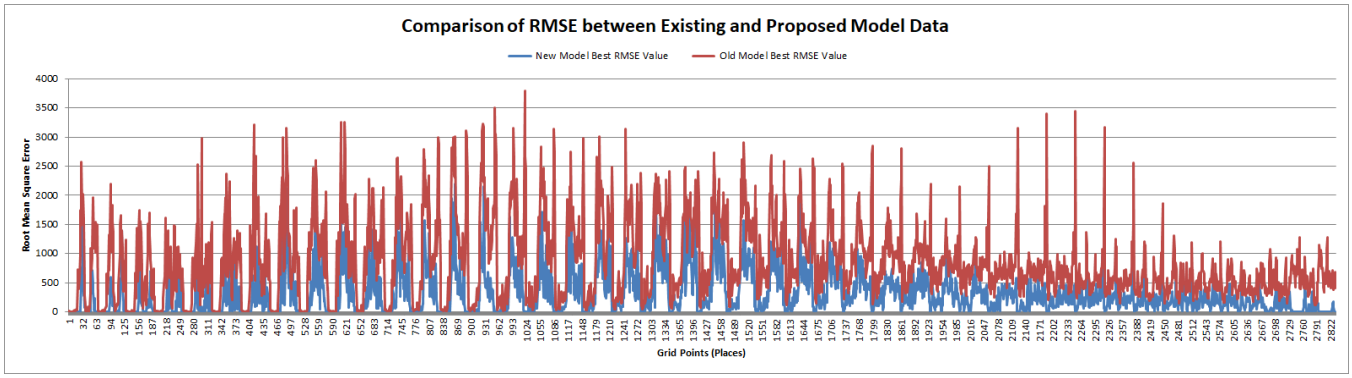

Fig. 6. Comparison of Root Mean Square Error (RMSE) between existing and proposed model data for each individual grid point.

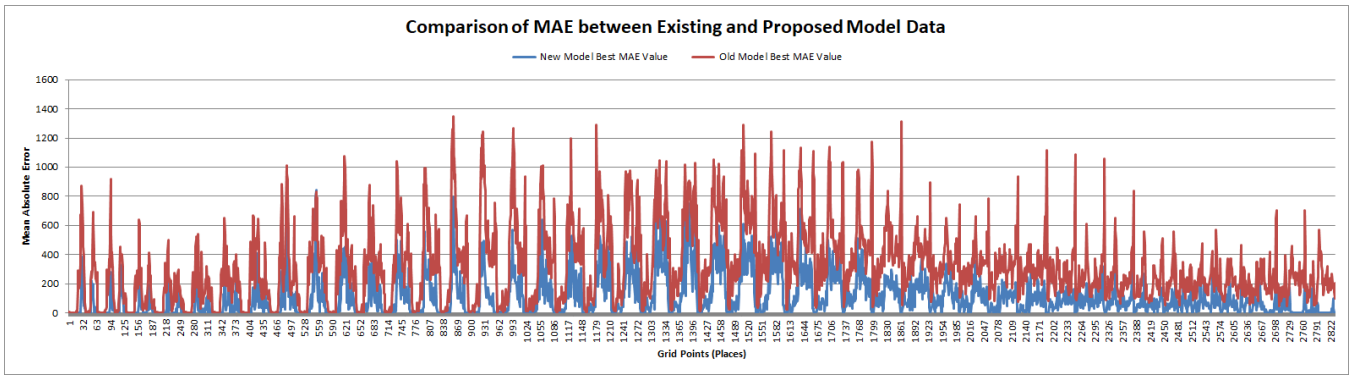

Fig. 7. Comparison of Mean Absolute Error (MAE) between existing and proposed model data for each individual grid point.

TABLE I
CHART OF ERROR RATE REDUCTION

Dataset	Error Measure	Old Average Error - New Average Error	Error Reduced	Percentage of Reduced Error
25x25 Summing Dataset	RMSE	846.0578 - 294.0196	552.0382	65.2483%
	MAE	313.5924 - 104.9345	208.6578	66.5379%
15x15 Summing Dataset	RMSE	224.3965 - 116.4469	107.9487	48.1064%
	MAE	84.3705 - 38.5208	45.8497	54.3432%
5x5 Summing Dataset	RMSE	30.0194 - 13.0161	17.0033	56.6410%
	MAE	11.9613 - 4.1441	7.8172	65.3543%

existing prediction models. On the other hand, for 15x15 summing dataset, 99.06% cases new model's MAE is better than any other existing prediction models. Again, 100.0% cases RMSE of new model is better than existing prediction models. Again, for 5x5 summing dataset, 99.84% cases new model's MAE is better than any other existing prediction models. Again, 100.0% cases RMSE of new model is better than existing prediction models.

Fig. 6 illustrates the comparison of Root Mean Square Error (RMSE) between the existing data and proposed model data for each individual grid point of 25x25 summing dataset. The green line which indicates the RMSE of new model, shows clearly that the new model reduced the error rate more than half. Again, Fig. 7 shows the comparison of Mean Absolute Error (MAE) where the new model reduced MAE more than half.

Table 1 shows the result of the new proposed model. In other word, it shows how much the new model was able to reduce the error rate. For 25x25 summing dataset, around 65% RMSE and 66% MAE have been reduced. Around 48% RMSE and 54% MAE reduction have been earned for 15x15 summing dataset. Again, for 5x5 summing dataset, the reduction of RMSE and MAE were around 56% and 65%.

Since, we used five different machine learning models (Linear Regression, K-Nearest Neighbors, Neural Network, Random forest, Support Vector Machine Regression), and picked the best model all the time, we found overall Linear Regression (selected as best model in 1050 places for 25x25 summing dataset, in 3143 places for 15x15 summing dataset, in 28873 places for 5x5 summing dataset) performed better than other four machine learning model. Again, K-Nearest Neighbors (selected as best model in 806 places for 25x25 summing dataset, in 2103 places for 15x15 summing dataset, in 15934 places for 5x5 summing dataset) and Random forest

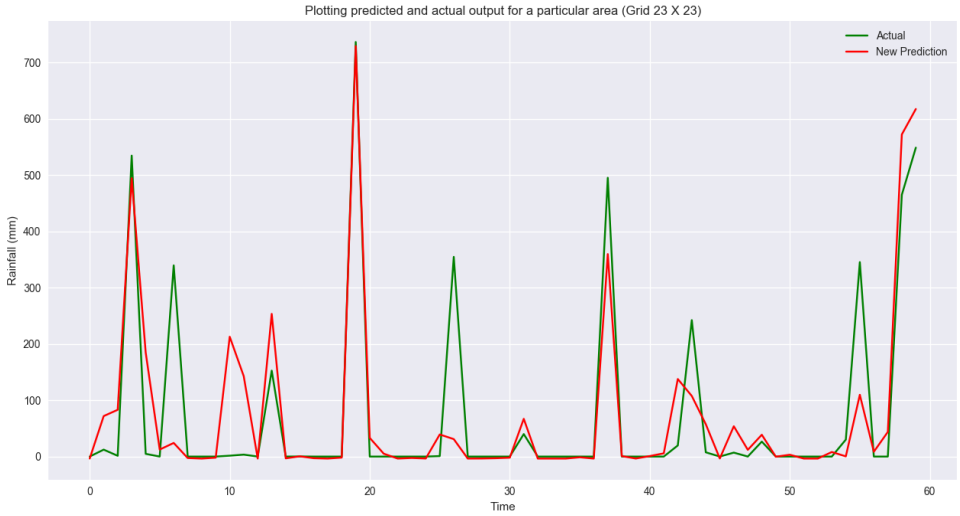

Fig. 8. Plotting predicted and actual output for a particular area (Grid 23X23).

(selected as best model in 647 places for 25x25 summing dataset, in 1775 places for 15x15 summing dataset, in 12986 places for 5x5 summing dataset) were also comparatively performed good.

Fig. 8. shows the actual and predicted result for some random testing data for grid point 23x23. This particular prediction was given by random forest. If we analyze this prediction graph, the model was able to predict almost all of the heavy rainfall, though there were some mis-prediction for some particular time.

V. CONCLUSIONS

Aggregated model what has been proposed in this paper is showing comparatively better result for most of the cases. But, reducing error rate more is necessary and we believe, the error rate could be reduced more. Non-machine learning approach (e.g. combining all the existing prediction model data by using some kind of voting technique) might be another option.

It is planned to go forward for further research to make a better model and try with more high resolution grid.

Since, researchers usually don't use other models data to develop new aggregated model, we believe, this research would open a new approach of improving our models in a different way.

ACKNOWLEDGMENT

The authors would like to thank the professor of Computer Science Department, New Mexico Tech, Dr. Hamdy Soliman for providing the access in his powerful machine for this research work.

REFERENCES

[1] Wei-Chiang Hong, Rainfall forecasting by technological machine learning models, Applied Mathematics and Computation, Volume 200, Issue 1, 2008, Pages 41-57, ISSN 0096-3003

[2] THOMPSON, J.C., 1950: A NUMERICAL METHOD FOR FORECASTING RAINFALL IN THE LOS ANGELES AREA. Mon. Wea. Rev., 78, 113124

[3] Hernndez E., Sanchez-Anguix V., Julian V., Palanca J., Duque N. (2016) Rainfall Prediction: A Deep Learning Approach. In: Martnez-lvarez F., Troncoso A., Quintin H., Corchado E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2016. Lecture Notes in Computer Science, vol 9648. Springer, Cham

[4] Beda Luitel, Gabriele Villarini, Gabriel A. Vecchi,Verification of the skill of numerical weather prediction models in forecasting rainfall from U.S. landfalling tropical cyclones, Journal of Hydrology, Volume 556, 2018, Pages 1026-1037, ISSN 0022-1694

[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14(Aug):23492353.

[7] Svante Wold, Kim Esbensen, Paul Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems,Volume 2, Issues 13, 1987, Pages 37-52, ISSN 0169-7439

[8] Herve Abdi and Lynne J. Williams, Principal component analysis, 2010 John Wiley and Sons, Inc. WIREs Comp Stat 2010 2 433459

[9] Trevor Hastie and Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning - Data Mining, Inference, and Prediction

[10] Steve R. Gunn, Support Vector Machines for Classification and Regression, UNIVERSITY OF SOUTHAMPTON, 10 May 1998

[11] Simon Haykin, Neural Networks: A Comprehensive Foundation (3rd Edition), 2007, 0131471392, Prentice-Hall, Inc.

[12] Leo Breiman, Random Forests, Mach. Learn., 0885-6125, 45, 1, 5-32, 2001

[13] Xin Yan and Xiao Gang Su. Linear Regression Analysis: Theory and Computing. World Scientific Publishing Co., Inc., River Edge, NJ, USA. 2009.

[14] Verleysen, Michel & Franois, Damien. (2005). The Curse of Dimensionality in Data Mining and Time Series Prediction. Lecture Notes in Computer Science. 3512. 758-770. 10.1007/11494669_93.

[15] Pedro Domingos. A few useful things to know about machine learning. Commun. ACM 55, 10 (October 2012), 78-87.