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`

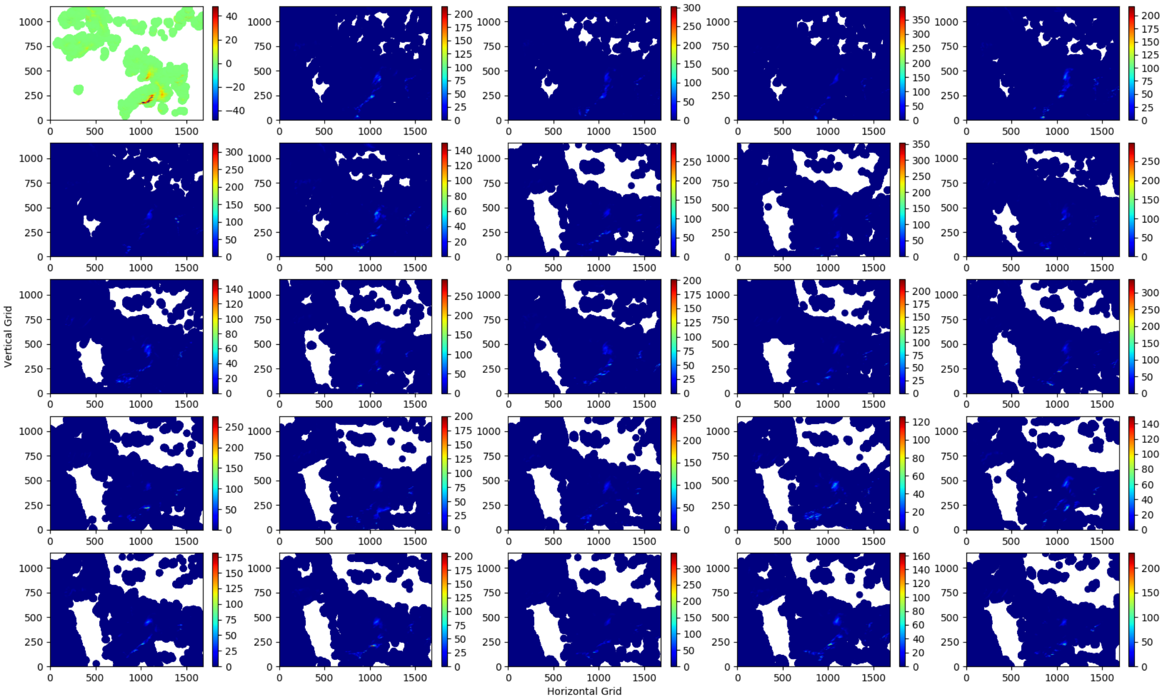

Fig. 1. Visualization of precipitation real verification data and prediction models' data. The first image (image in the top left corner) is the real verification precipitation data and rest of the 24 images are visualizing precipitation prediction data from 24 prediction models.

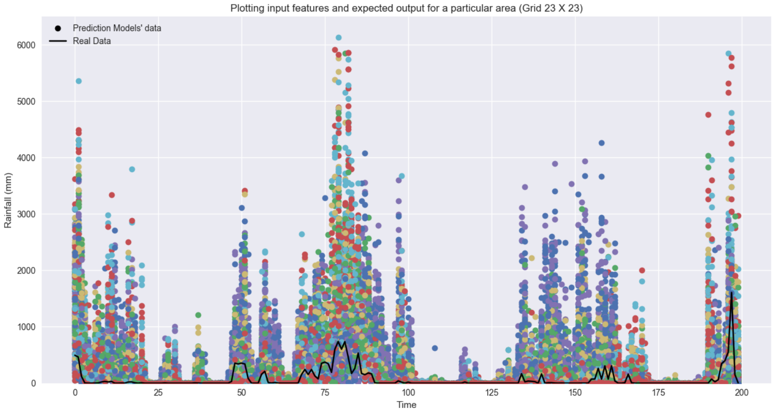

Fig. 2. Plotting input features and real verification data for a particular area which is position 23 X 23 in the grid.

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.

Fig. 1. illustrates the comparison between 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 predictions of existing data comparing to the real data. It illustrates the rate of over-predicting. All the prediction models we have, provides high error rates. 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 didn't have data for a large amount of time. Again, some days didn't have data for lots of prediction models. The data that has been used to develop a new aggregated model has 24 prediction models for 20 days in total and 10 different times for each day.

A. Resolution selection

We have the prediction data for 1155X1683 different geographical coordinates for all over the USA. Since, every coordinates cover little more than 1 mile square, the initial goal was to find out a stable resolution for comparatively big 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 after getting a good for a 25X25 area, we tried to make the resolution higher. The next resolution grid size we used was 15X15.

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.

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. 3. 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 error 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. To find out the best features, a feature selection technique has been used which is discussed in the next section.

D. Univariate feature selection

Univariate linear regression shows the relationship between dependent and independent variable. [6] Univariate feature selection is a linear regression technique to test individual importance of each independent variable. [5]

After using the univariate feature selection technique, most effective 50 features have been kept out of 216 features. Fig. 4. shows the over predictions of the selected 50 features. Next target is to use feature extraction technique and then fit these input dataset into machine learning model.

E. Dimensionality Reduction

Since, the input dataset has only 200 samples, so it is important to reduce the dimension according to curse of dimensionality. Again, over-fitting should be an important concern. Principal Component Analysis (PCA) has been used to reduce dimension of our input vector. Principal component analysis convert the high dimensional data into lower dimension.The goal of doing PCA is to summarize the data into limited number of principal components. [7] [8]

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 assigns weight from neighbors based on the contribution of neighbors. K-nearest neighbors algorithm works on both classification and regression problem. [9]

2) Support Vector Machine: Support vector machine is a supervised learning algorithm which works on both classification and regression problem. Basically, it build a hyperplanes in a high dimensional space. [10]

3) Neural Network: Neural network is based on multiple artificial neurons. Each artificial neuron belongs to [11]

4) Random Forest: Random forest is one of the ensemble method which can work on both classification and regression. It constructs multiple decision tree in the training phase and outputting the class which is a mode of the classes for the categorical approach. [12]

5) Linear Regression: Linear regression creates a model that represent the relationship between the dependent variable and one or more independent variable. [13]

IV. Experimental Result

We got a satisfactory result with low resolution grid (25X25). Then, the resolution has been increased and still the result was promising. Fig. 5. shows the actual and predicted result for some random testing data. This particular prediction was given by random forest, though both linear regression and k-nearest neighbors are also giving good results for some particular places.

TABLE I

Results illustrate how many cases the new model has lower error rate

Grid	MAE	RMSE
25X25	98.44% cases	100.0% cases
15X15	99.06% cases	100.0% cases

For 25X25 grid, 98.44% cases new model's MAE is better than any other existing prediction models. Again, 100.0%

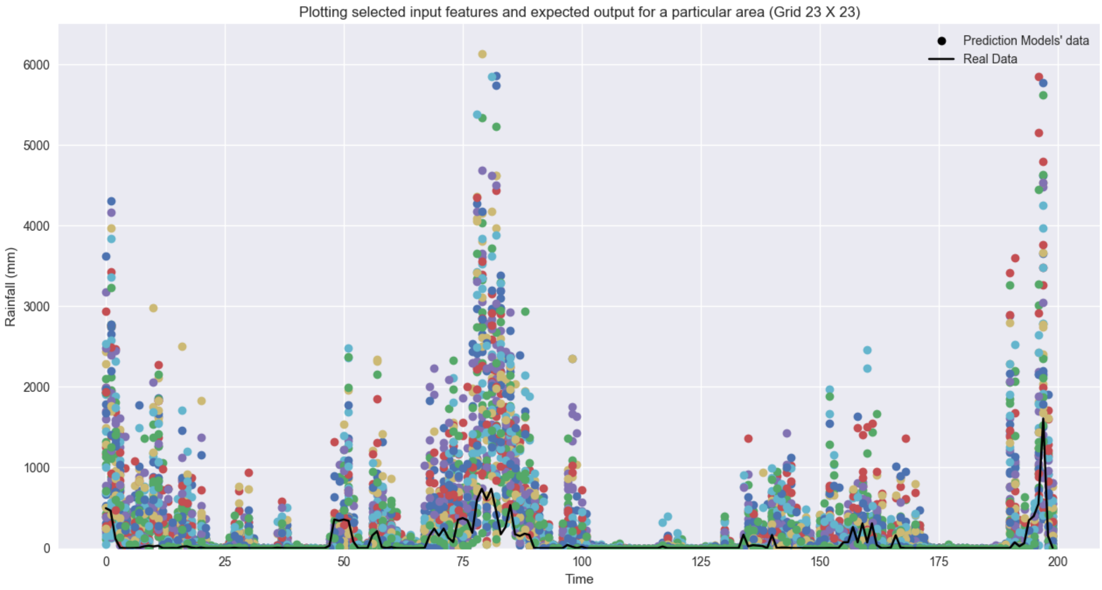

Fig. 3. Visualizing selected 50 features and the target values

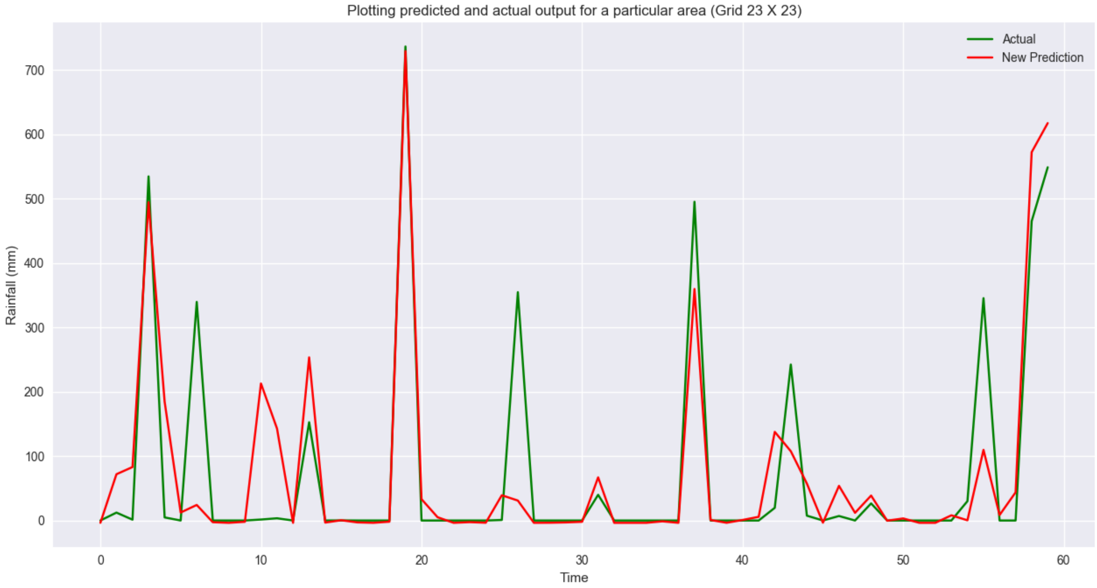

Fig. 4. Plotting predicted and actual output for a particular area (Grid 23X23).

cases RMSE of new model is better than existing prediction models (shows on Table 1).

On the other hand, for 15X15 grid, 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 (shows on Table 1).

Fig. 6. shows the comparison between MAE of new model and the best existing data. The new prediction model has highest MAE value little more than 800, while the existing models' best MAE has value more than 1200. Again, most of the cases the new model's MAE is less than 150.

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,

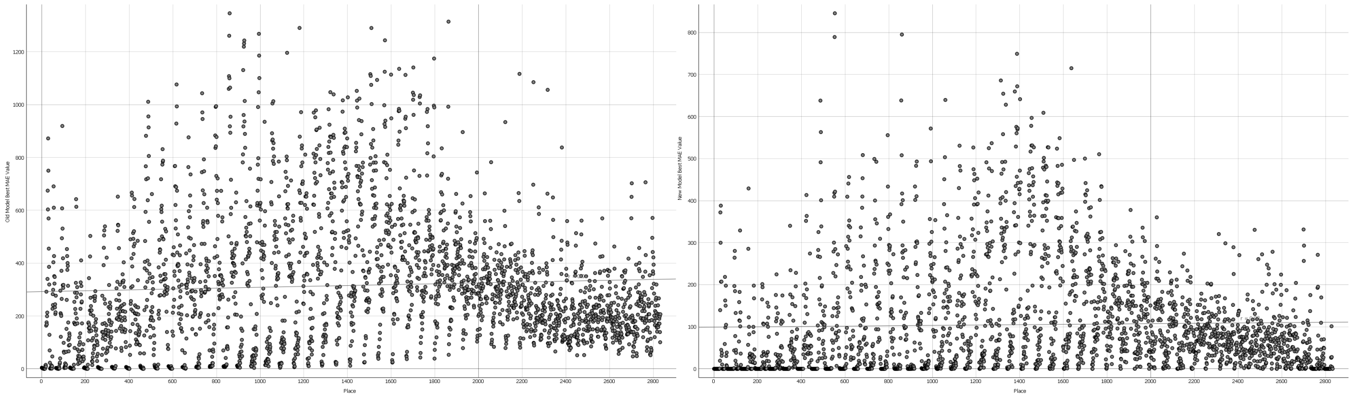

Fig. 5. Comparison of MAE of the best existing models and the new model.

the error rate could be reduced more. Non-machine learning approach by combining all the existing prediction model data 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] Schneider, Astrid, Gerhard Hommel, and Maria Blettner. Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. Deutsches rzteblatt International 107.44 (2010): 776782. PMC. Web. 7 May 2018.

[6] https://docs.orange.biolab.si/3/data-mining-library/reference/preprocess.html

[7] Jake Lever, Martin Krzywinski, Naomi Altman, Principal component analysis, Nature Methods, 2017/06/29/online, 14, 641, Nature Publishing Group, a division of Macmillan Publishers Limited.

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

[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967.

[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] Schneider, Astrid, Gerhard Hommel, and Maria Blettner. Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. Deutsches rzteblatt International 107.44 (2010): 776782. PMC. Web. 8 May 2018.