Aggregating Ensemble Weather Predictions for Rainfall Predictions

***Abstract*—Atmospheric science extensively uses conceptual models that simulating weather. Traditionally, conceptual models are dynamical systems represented as differential equations involving 'important' variables. A meteorologist faces the challenge of choosing from the (often conflicting) predictions of different conceptual models.**

In this paper, we explore applying Machine Learning techniques to aggregate conceptual weather models of rainfall. Specifically, we explore the question 'is it possible to form an aggregate model of existing conceptual models for predicting rainfall based only on the outputs of these models (that is more accurate at predicting rainfall)?'.

Our secondary goal is to advance atmospheric science. We explore the correlation of predictive power of different conceptual models. This is a first step in categorizing the predictive power of the components (variables and parameters) of various models as a function of geographical location, time, etc. This information, in turn, helps in developing better conceptual models as well as making better aggregated predictions.

We aimed to build an aggregate model for each geographical grid point, and at more than half the geographical locations, we were able to build ones that outperform existing conceptual models. We found that locations where the aggregate model was better than the existing ones were such that the best existing model had low errors, i.e., where the existing models are good, the aggregate model is better. We also discovered that locations that are geographically near each other tend to use the same group of conceptual models to build the aggregate one. We plan to analyze this to discover the commonalities in the conceptual models used with an aim to determine which model parameters are best suited for which area.

***Index Terms*—machine learning, weather forecasting, rainfall prediction, post-processing, aggregation.**

I. Introduction

Weather is a complicated, chaotic system encompassing many variables. The qualitative and quantitative behavior of this system is hard to predict. This is evidenced by the existence of many models outputting conflicting, if not contradictory predictions regarding the same weather phenomenon. There have been two main approaches to predict weather events - using models based on atmospheric science which are concept and science driven, and using machine-learning based models which are data driven. These have largely been independent of each other in their methodology and use.

Atmospheric Science uses conceptual models to simulate the atmosphere and the weather events therein. Weather is modeled as a dynamical system, and represented as differential equations between meteorological variables; these differential equations are integrated (numerically) to simulate and predict weather events. The different models differ in the variables they consider important, the spatial and temporal resolution of the dynamical systems they simulate, meta-parameters etc.

Advancement in such modelling would involve improvements in the concepts such as changes in the relative importance of variables with change in the underlying geography. A short-term goal would be to assist meteorologists considerate the disparate predictions of the many models into a single prediction that can then be used to issue severe weather watches and warnings.

Machine learning models use existing data to construct a function from features that are relevant to weather events of interest (which may be quantifiable). These models are obtained using various algorithms, all of which try to optimize the function by maximizing its accuracy in predicting. There have been many such developments.

Traditional statistical analysis techniques have been used for precipitation prediction. J. C. Thompson [1] 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 [2] 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 Numeric 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 new models have been proposed that map meteorological variables such as temperature, wind speed, humidity etc. to desired weather events such as precipitation. Most of the cases, the error rates for the prediction data of those new models are high, e.g., [2] [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 duration, but not for longer duration. [4]

A. Larger vision and present work

While machine learning based models can outperform atmospheric science based models, they do not help atmospheric science. In particular, many of the machine learning models are black-boxes, and even in the case where they are not, the models are very complicated to help advance the scientists' understanding of the phenomena being studied. This is especially the case where there are several existing models modeling the same phenomenon, and ML model becomes yet another model, albeit (possibly) better at prediction. Here, we present results from a project that forms part of a larger goal to use machine learning algorithms to help advance scientific understanding.

Model Validation for conceptual models is an important area of research in meteorology. The aim of this field is to determine which conceptual model works best under which set of circumstances. As part of our larger project, we wish to use machine learning to address this question. Here, we present the first steps in this.

Our main goal in this work is to explore if it is possible to assist a meteorologist consolidate conflicting outputs from disparate conceptual models of rainfall. At the onset, it is unclear whether this is possible because we do not want the ML models to have access to anything but the outputs of these conceptual models. One reason for doing this is to simulate a scenario similar to one faced by a meteorologist faced with this task in a strict sense; another reason is that many ML algorithms are powerful enough that given sufficient features, they can create a new model from scratch, which would defeat our goal. (There are restrictions to ML algorithms to prevent this, which we plan to explore in future work.) Because of this design decision, ML algorithms do not have the features that may influence different conceptual models' relative performance. For instance, considering conceptual models that predict rainfall it may be the case that one model may have more accuracy than another when the temperature is high; but since the model output is only rainfall, this information is unknown to the ML algorithm trying to aggregate them. If ML models are successful at this, it would be a proof of principle for the claim that ML can help in advancing basic science. Even a partial success would give us information regarding the conditions under which this is possible.

Our secondary goal (that extends beyond this work) was to explore whether machine learning can help advance atmospheric science. In particular, we want to see whether we can find geographical features (coastal, mountainous etc) within which the same set of conceptual models perform well; Finding the commonalities in these conceptual models with the help of a domain scientist might shed light about variable relevance and parameter tuning based on geography.

We explored these questions using models that predict rainfall. We were able to build an aggregate model that outperform existing conceptual models. We found that locations where the aggregate model was better than the existing ones were such that the best existing model had low errors, i.e., where the existing models are good, the aggregate model is better. We also discovered that locations that are geographically near each other tend to use the same group of conceptual models to build the aggregate one. We plan to analyze this to discover the commonalities in the conceptual models used with an aim to determine which model parameters are best suited for which area.

II. Data Science Research

The Weather Research and Forecasting Model (WRF) provides convection-allowing forecasting capabilities through a Numerical Weather Prediction framework that allows the construction of a multi-model ensemble using various combinations of surface, turbulence and micro-physics options as well as permutations in the initial conditions [5]. The Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU) has been running WRF based weather forecasts over the contiguous United States (CONUS) on NSF eXtreme Science and Engineering Digital Environments (XSEDE) [6] resources as part of a collaboration with the NOAA Storm Prediction Center and the National Severe Storms Laboratory. The use of high resolution convection-allowing grid scales of 3-km or finer allows the direct numerical modeling of convective storm processes that otherwise would have to be estimated via parameterizations. The computations involved in the Spring Forecasting Experiment are enormous, as are the simulation outputs; in the Spring 2016 experiment, for example, simulations consumed 16.27 million XSEDE service units (computing hours) on the Stampede cluster at the Texas Advanced Computing Center. These ensembles are evaluated as part of the NOAA Hazardous Weather Testbed activity in real time [7]. The Hazardous Weather Testbeds Spring Forecasting Experiment is the largest coordinated effort in the country to test and improve the use of ensemble forecast data from multiple simulations of WRF model.

Beyond their evaluation in real time for improving operational forecasts, the numerical weather prediction model outputs are rich sources for data exploration. For example, these data have been mined using Random Forests for algorithms to improve hail forecasts [8] in a project that requires many software steps to train and validate the algorithm involving oversight by data mining experts and meteorologists. Separately, Brewster et al. [9] have been using NCARs Visualization and Analysis Platform for Ocean, Atmosphere, and Solar

Researchers (VAPOR) program to visualize model variables and combinations of model variables in the 3-km resolution weather simulation output files as a way to try to visualize and understand how the evolution of different 3-dimensional features in the model output. These visualizations illustrate derived features that exist in the data that must be extracted using phenomenological models created from the insights of atmospheric scientists with many years of expertise.

These simulation data sets are too large to explore fully using current methods and infrastructure. Furthermore, human expertise is required for selecting 2D variables extracted from the 3D data set for real-time use and later verification and study. This expertise requires years of experience from past experiments as well as professional training and thus combines intuition, experience, and concrete physical models. Thus we have two challenges. First, there are a limited number of highly experienced scientists capable of understanding the data. Second, even with years of experience, it is not possible for an unassisted scientist to identify and examine all potentially interesting features in the data. It is evident that automating the feature detection that underlies the visualizations can aid understanding and serve as a source of new variables to explore with machine learning.

In this paper we explore use of machine learning techniques to analyze, verify, aggregate and post process large number of ensemble weather simulations.

III. Methodology

A. Data

1) Observed rainfall data: The weather simulation data sets create both direct and derived data on spatial-temporal grid points covering contiguous United States. These data sets are useful not only for real-time forecasting, but also for retrospective studies. These forecast domain have data grid points at 3km grid spacing with 1683 grid points in east-west direction, 1155 in north-south direction and 53 vertical levels. (The vertical levels are not used in the rainfall dataset). The observed data are only available over land. The dataset also contained 20 different days in the season of spring, starting from was 04/19/2016 and ending at 05/16/2016 and for each day we had 10 different times.

We present here some basic statistical information about the real data, (i.e., observed rainfall). In Figure 1, 1(a) plots the average value of real precipitation for each geographical grid point. Each individual grid point contains the average precipitation value of that place over all days and times. Figure 1(b) shows the standard deviation of real precipitation for each grid point. There are a large number of grid points where the the standard deviation is pretty high. (plotted this against geographical coordinates)

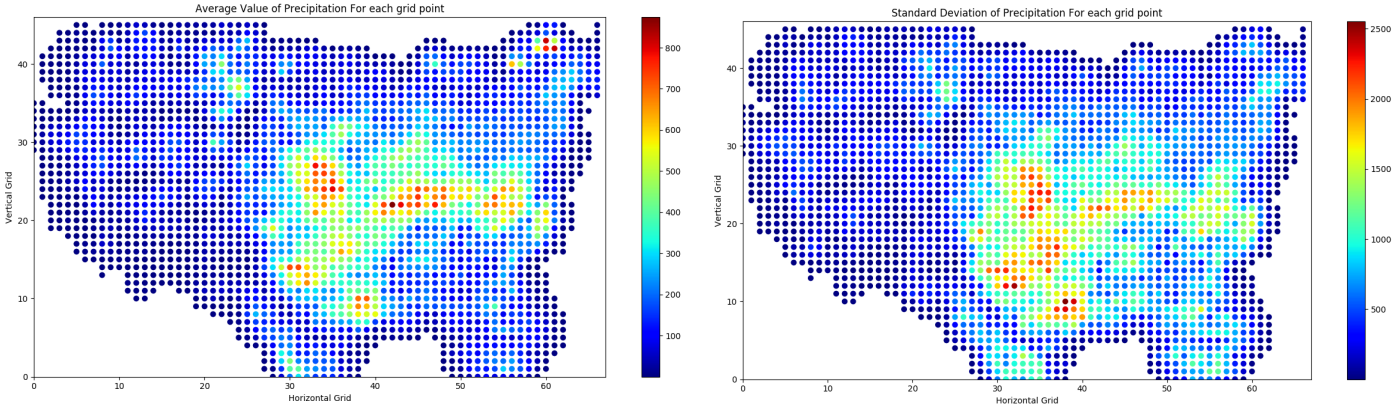
2) Existing Conceptual Weather Models of Rainfall predictions: Forecasts upto 72 hours are outputted at 6 min time intervals. These WRF simulation output data is stored in compressed NetCDF format in split-file form. A wide range of scientific data types such as the forecast outputs use NetCDF due to their self-describing data formats, which are parsable by both programs and scientists. Our dataset contained 24 conceptual models that predict rainfall 3 hours into the future.

Our main goal is to simulate a meteorologist who aggregates the outputs of conceptual models. So, our datasets just contained the model outputs, and each model output contains information about the time and spatial grid point of the prediction, and the rainfall amount predicted by that model. Traditional machine learning approach to this problem would be to create an ML model that has access to weather parameters which may affect the weather models' performance. However, this introduces the risk of creating a new ML-based model of predicting rainfall instead of aggregating the old ones. Instead, we explore ML models which can only utilize output of existing weather simulations.
Our design decisions were influenced by several secondary goals. One such important secondary goal was to contribute to the atmospheric modeling. Here we present statistical analysis of these models, and our feature selection process, and the ways we deviate from a typical machine learning approach.

a) Statistical Analysis and Feature Selection: There are many ensemble modelling techniques but these rely on creating many random (ML) models, each one having very good performance under different circumstances, and cleverly aggregating them. Here, we are given specific (weather) models we wish to aggregate. The performance of these weather models may differ based on atmospheric parameters such as temperature, which we do not have access to. At first glance, this problem seems one of simple post processing, but the following feature analysis shows that there are enough features that can influence the aggregation. Thus, posed as an ML problem, creating an aggregate model is non trivial, but it is unclear on the onset whether the outputs have enough information for the aggregate model to outperform existing ones.

Features that can be extracted purely from the output of models which may impact the aggregation. These include

- Geographical Location: Latitude and Longitude of the grid points of prediction is available. Using this, the following information can be extracted statically.
  - geographical category, such as mountain, desert, plains,
  - rural, urban, forest areas
  - distance from large water bodies
  - altitude
  - prone to severe weather (of different categories. e.g., tornadoes, hurricanes, flash floods, heat waves)
- Time: Date and time of prediction is available. Using this, the following information can be extracted.
  - time of day (this will affect the temperature)
  - time within a season (this affects temperature, and the possibility of severe weather events)
  - time from the last severe weather event
  - time into the current severe weather event if applicable



(a) Average Precipitation for each grid point

(b) Standard Deviation of Precipitation for each grid point

Fig. 1. Analyzing weather simulation outputs for rainfall

- Range of output predictions from the model: This information is directly available. The reasoning behind this is that when the observed rainfall amounts may determine the performance of model predictions. This is complicated by the fact the the parameters of some of the models are tuned to predict severe weather.
- Severe weather events: This information is not directly available from the output, but using the predicted output values along with recent past predictions, this can be determined.
  - The distance and time from the focus of the severe weather event
  - The direction of movement of the severe weather event
  - The consensus or disagreement of the severe weather events' focus and direction among various models

We will consider the features in some detail here.

3) Geographical information: Basic statistical analysis of existing models showed that absolute spatial location was the most influential factor that explained the difference in model performance. The performance of the models is plotted in Figure 2 which show the root mean squared error of the models (for rainfall summed up temporally). [1] The following plots are over the continental US. The models included predictions over the sea, but since there are no observations over the sea, this gives an error. Here we do not remove this area in order to show the model behaviour. (We do remove this area in subsequent sections).

The plots show that most of the models perform well in areas of low to moderate rainfall, and have more errors on locations of high rainfall, or rather, high standard deviation. This is not surprising. However, careful analysis of the the error plot shows that there are slight variations in the regions of high error for each model, and even within the area with high rainfall, this pattern remains. The interesting thing is that there are patches of areas with similar error rates - i.e., there are contours of this plot. This information shows that adding geographical information of the locations may provide useful information about model accuracy. A related question which we delve into later in this paper is whether the a subset of models performing well in one geographical area, say desert, will do so in other deserts as well.

Since geography had a big role to play in accuracy of models, and predictions are made at each geographical location, as well as the aggregate results made by humans, we decided to develop aggregate models for each individual location. The goal was to analyze (in future work) the aggregation to determine individual model influences based on geography. This also led us to try coarser resolutions that is available in the data, since that would create a model for a local geography. In general the preliminary results degraded slightly with coarser resolution, but since local geography typically covers a larger area. In the following, we consolidate 25*25 grid points into a single value at its center - we call this a spatial unit. Data might be too small for temporal factors' significance to show up; Model performance was worse for certain real outputs, but this was true across all models, and so this did not modify aggregation parameters much.

4) Interval of observed rainfall: Here, we explored whether the models' accuracy depended on the observed rainfall, i.e., are there some models which perform better predicting low rainfall than high rainfall? To further understand the performance of weather simulations in predicting rainfall, we plotted the average precipitation prediction of a simulation for each real precipitation value as illustrated in Figure 6 in the appendix. While there seemed to be correlation between observed rainfall and error rates, we had to take into account the fact that in the geographical regions with high rainfall, we also had high variability of rainfall. Further, the correlation between observed rainfall and error cannot be directly used since we will not have access to the observed rainfall as part of the conceptual models' outputs. And when we used this feature along with geography, this effect diminished, since there is a strong correlation between geographical location and amount of rainfall.

[1] Other error measure plots are available in the appendix.

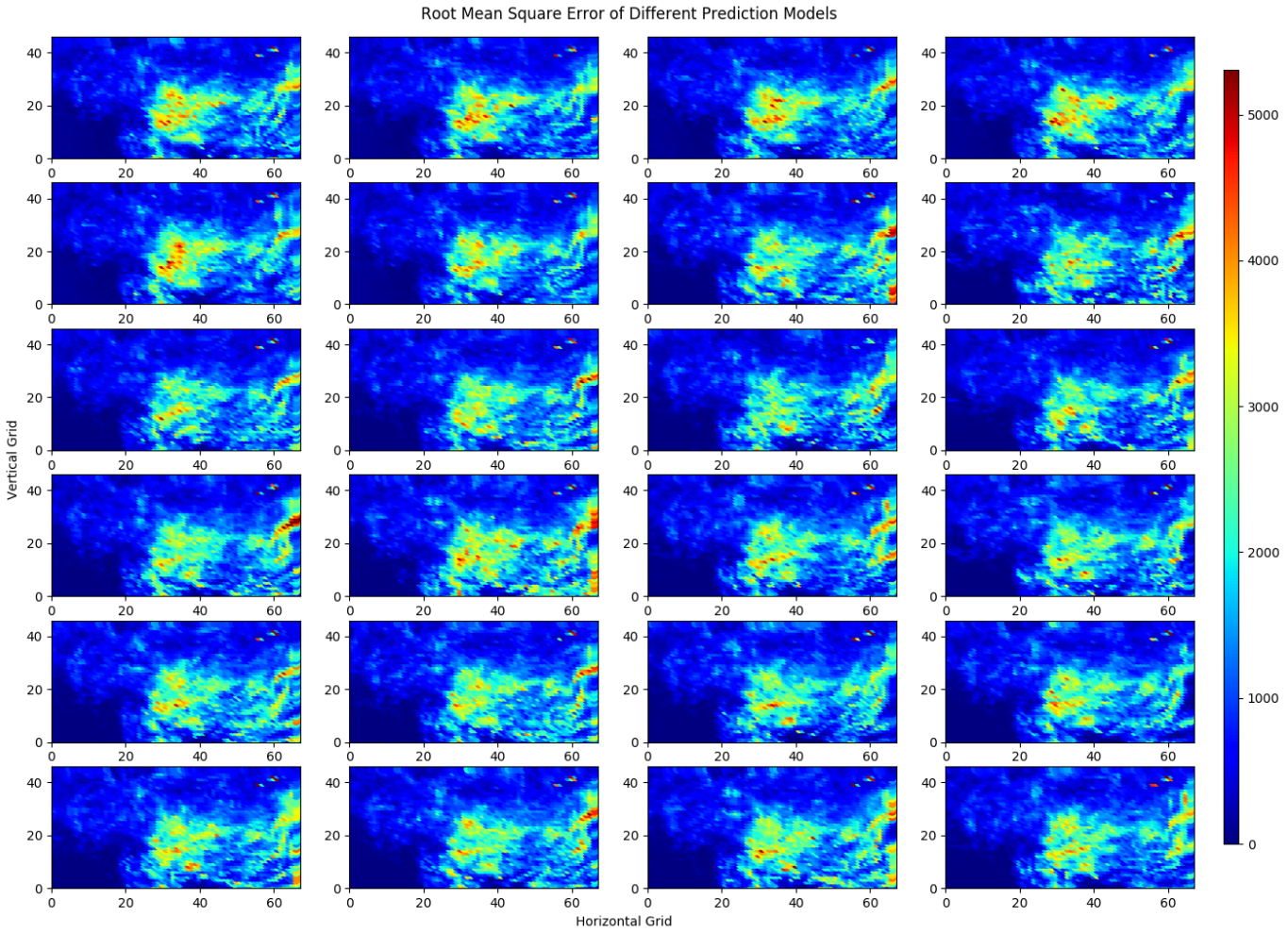


Fig. 2. Errors of Different Prediction Models

We did not use this feature explicitly; however several ML methods, especially methods like KNN indirectly use this information.

5) Temporal and other features: The current data set is too small for temporal features and features involving severe weather to be tested. This is mainly because for temporal aspects to be taken into account, we would need to consider time series analysis of the predictions, and this is especially important during severe weather events which cause high rainfall. However, the number of severe weather events occurring in a season is limited to verify this. We plan to expand the data set in our future work.

B. ML models

The atmospheric models we are trying to aggregate are based on differential equations simulating a dynamical system. Since the behaviour of linear combinations of dynamical system is qualitatively different from non-linear combinations, linear combination of weather models are preferable. It would be interesting to know which geographical regions are modelled well enough by existing models that linear combination is a sufficient aggregation method.

So, for each spatial location, we developed many machine learning models to aggregate existing models, and chose the best one (based on test set error, and we report error on unseen validation set here).

1) Problem Statement: For each geographical unit (25*25 grid points), we develop an ML based aggregate model (that may use different ML algorithm for different spatial location). The feature set included 3*3 units centered around the value being prediction from each model - thus we begin with 216 features. We used statistical methods such as correlation, information gain etc. to pick 50 features out this, and used the first 5 components of PCA. The machine learning models included regression (linear and polynomial), KNN (7 neighborhood), random forest, SVM (Radial Basis Function kernel), and feed forward neural networks (We did not have enough data to divide the segment into temporal chunks to train RNNs. We hope to do this in future). We also included naive weighted average weighted by accuracy of prediction.

TABLE I
AVERAGE ERRORS OF CONCEPTUAL AND AGGREGATE MODELS (AVERAGED OVER LOCATIONS)

Description	RMSE
Average error of the best conceptual models (averaged over locations)	347.447
Average error of the best aggregate model (averaged over locations)	334.294

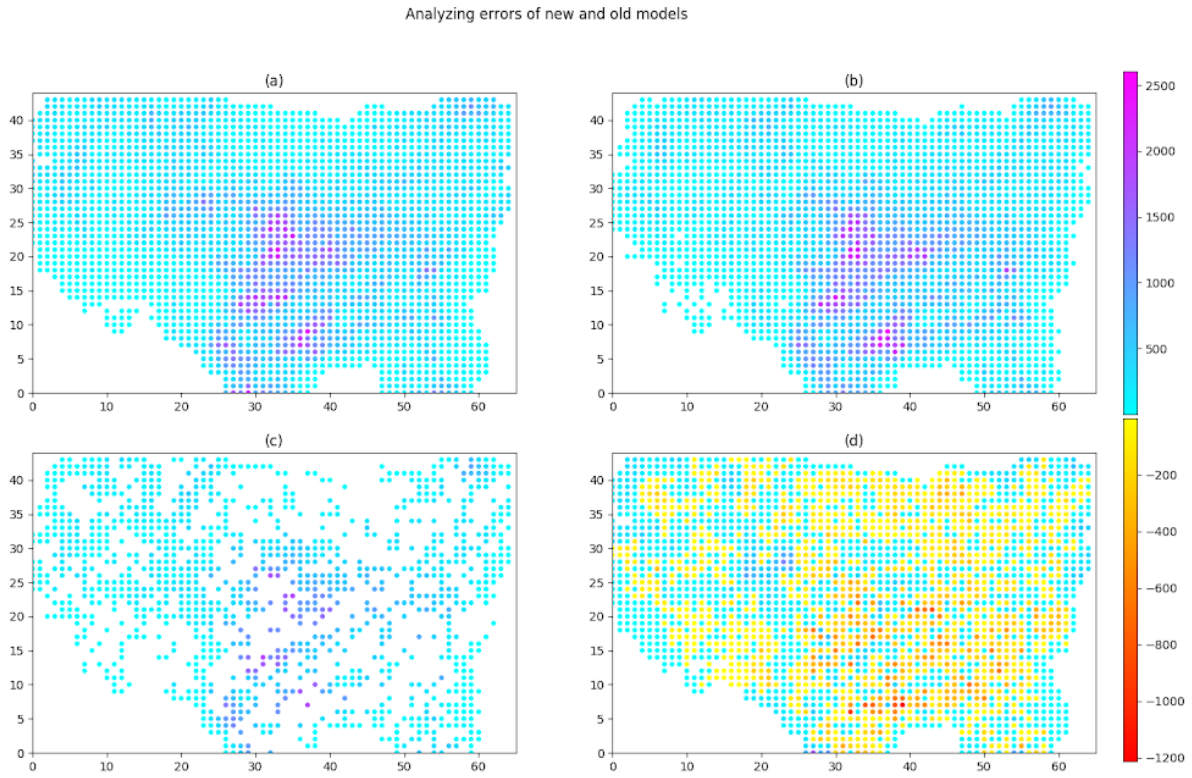


Fig. 3. Analyzing errors of aggregate and conceptual models

TABLE II
AVERAGE ERRORS OF NEW AND OLD MODELS BASED ON THE PERFORMANCE OF NEW MODEL

Description	RMSE
Average error of the best old model at locations where new model did better	241.074
Average error of the best new model at locations where new model did better	168.798
Average error of the best old model at locations where new model didn't do better	338.904
Average error of the best new model at locations where new model didn't do better	435.408

IV. CONCLUSION

A. Results and Discussion

In over half of the geographical location, the machine learning methods were able to successfully create an aggregate model that outperformed existing models. Table I lists the average error (averaged over locations) of conceptual and aggregate models. The Root Mean Square Error has been reduced by 13 which is almost 4%.

Figure 3 gives the error of the models old and new by geographical location. In all the plots, we give the root mean squared error. Figure 3 (a) gives the error of the best conceptual models at each location; (b) gives the error of the best aggregate model at each location; (c) gives the error of the best aggregate model at locations where the aggregate model outperformed conceptual ones; (d) gives the difference in the error of the best aggregate model and the best conceptual model at each location (Positive values indicate the places where new model is better, while negative values indicate where the old model is better.)

However, this result, on analysis yields surprising insights. Comparing the Figure 3 (a) and (c) (or (d)), we see that by and large, at the geographical locations where the existing models were performing well, the aggregation was successful,

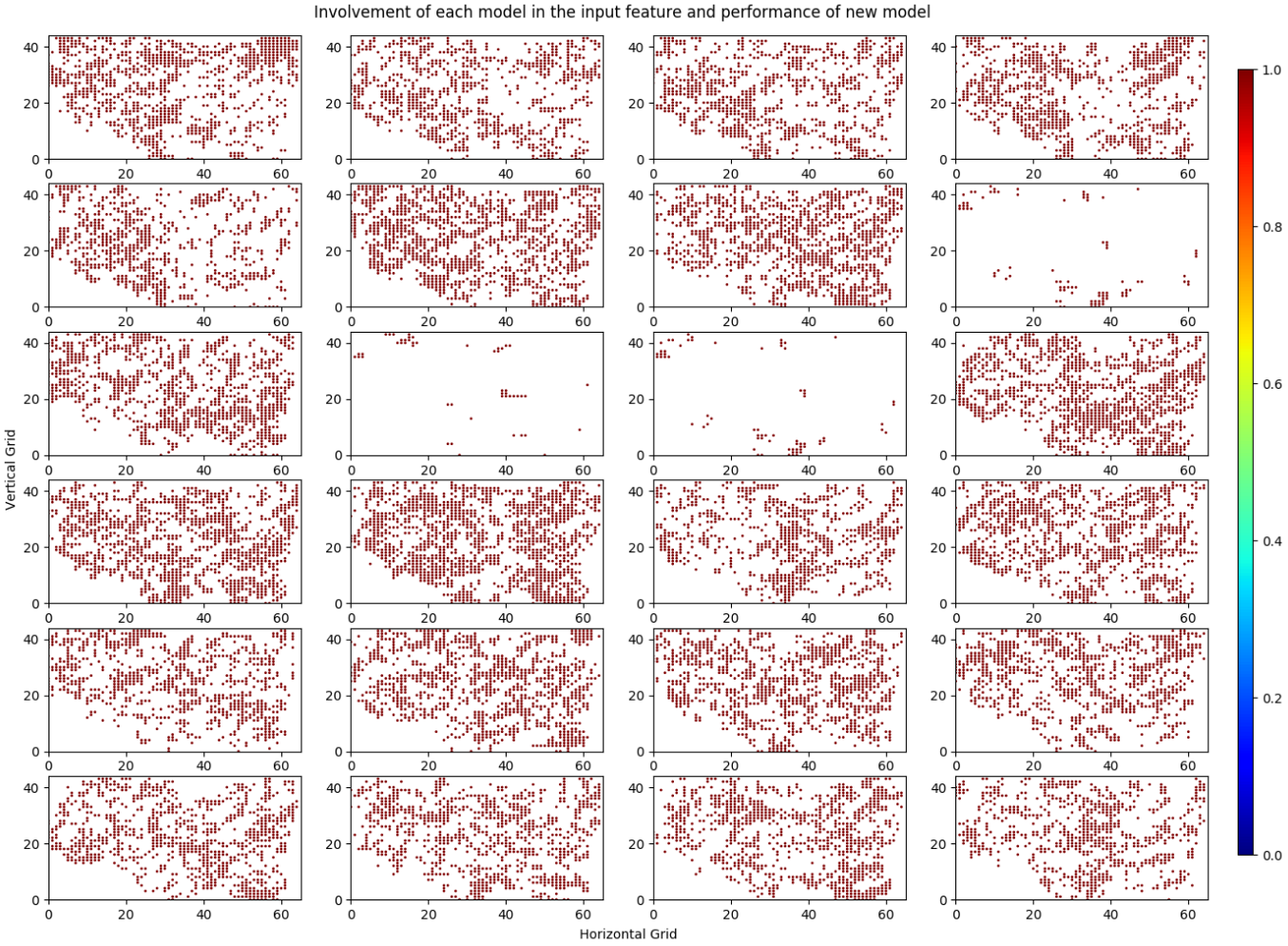


Fig. 4. Involvement of each model in the input feature and performance of new model

i.e., outperformed the existing models, (and the converse holds at places where the existing models perform poorly). This hypothesis can be verified as follows: divide the locations into those where the new model did best and those where some old model did best; then compare the difference between the error rates of the best old and new models at locations where new models perform best to the difference in the error rates at these two groups of locations. This analysis validated our hypothesis, and the results are given in Table II. At locations where existing models fail, at many places all the models make similar mistakes (i.e. over-predict or under-predict), and to figure out when they do so requires more information that is available in their outputs. At these locations, we need either more information that the output of the models, or we might want to develop a machine learning model that is data driven and replaces existing models. [2]

[2]The table giving the breakup of particular machine learning model's performance is available in the appendix.

The main question we set to answer was whether outputs of existing models is enough to create an aggregate model that is better than existing ones. The results were that it is possible a little over half the time, and corresponded to the places where the existing conceptual models were good.

The secondary goal we had was to contribute to atmospheric modeling. In particular we were interested in finding out if certain models performed well in certain geographical locations. To this end, we analyzed which models contributed positively to creating an aggregate model that was better than all existing models. In particular, we wanted to see if within a geographical area, the same models were most influential in creating an aggregate model which is plotted in Figure 4.

Traditional cluster analysis methods as well as image analysis such as connected component labeling and contour plots do not work well - the clusters and contours picked up are narrower than a visual analysis of the plots leads us to believe. In Figure 4, we could see that there was swatches of areas where a subset of model performs well. For instance, consider the north eastern part of the continental US. In several of

the models, there is NE-SW diagonal patch; some models are included here and some are excluded. We would like to extract this as a patch where different subsets of models behave in different ways. However, simple cluster analysis does not extract this clearly because there are other models that do well or badly in this patch, but only on part of the patch - many cover only the central diagonal line from these diagonal patches. So, we did a basic statistical analysis: At each spatial coordinate (x, y), let S be the bit-string with a bit dedicated to every existing atmospheric model and that has a 1 at bit i if atmospheric model i influences the aggregate model at (x, y). We plotted in Figure 5 the hamming distance of such strings against the distance that separated them. The following plot shows that at geographically close locations, the same models influence the aggregate model.

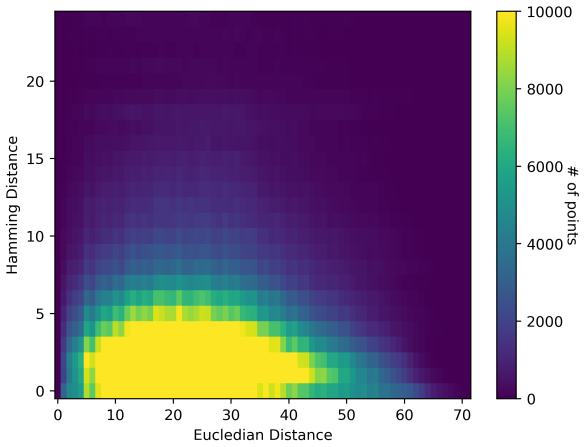


Fig. 5. hamming distance of the string S i have described there against Euclidean distance - this shows that grid points near each other have a lot of overlap in the models that contribute to the best models

We also plotted the geographical locations at which linear methods of aggregation performed best, and this again reinforces the idea that geography plays a major role in model influence and performance. The plot is attached in the appendix (Figure 12). We also noted several models that had very high influence on very small areas but had no influence outside these areas. We are planning on working with the domain experts to explain this.

B. Future Work

A very direct extension of this work is to train machine learning models in locations where the current models do not perform well. These ML models will include RNNs and others that capture the temporal nature of the data, and will include as features the weather data that is used in building atmospheric models. We also want to continue in other directions that would contribute to a better understanding of atmospheric modeling. In particular, we want to extract the models that do well on a geographical area, using a geographical map of US that has information about micro climates. From this, with the help of a domain expert, we plan to extract the properties that are common to models that perform well at a given geographical location. There are also important questions to explore that involve the four dimensional tracking of a storm.

ACKNOWLEDGMENT

Authors would like to thank Keith Brewster from University of Oklahoma for providing the data and guiding the authors about the problem.

REFERENCES

[1] J. C. THOMPSON, "A numerical method for forecasting rainfall in the los angeles area," *Monthly Weather Review*, vol. 78, no. 7, pp. 113–124, 1950. [Online]. Available: https://doi.org/10.1175/1520-0493(1950)078¡0113:ANMFFR¿2.0.CO;2

[2] W.-C. Hong, "Rainfall forecasting by technological machine learning models," *Applied Mathematics and Computation*, vol. 200, no. 1, pp. 41 – 57, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0096300307010843

[3] E. Hernandez, V. Sanchez-Anguix, V. Julian, J. Palanca, and N. Duque, *Rainfall Prediction: A Deep Learning Approach*. Austria: Springer Verlag, 2016, pp. 151–162.

[4] B. Luitel, G. Villarini, and G. A. Vecchi, "Verification of the skill of numerical weather prediction models in forecasting rainfall from u.s. landfalling tropical cyclones," *Journal of Hydrology*, vol. 556, pp. 1026 – 1037, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022169416305704

[5] J. Done, C. A. Davis, and M. Weisman, "The next generation of nwp: Explicit forecasts of convection using the weather research and forecasting (wrf) model," *Atmospheric Science Letters*, vol. 5, no. 6, pp. 110–117, 2004.

[6] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson *et al.*, "Xsede: accelerating scientific discovery," *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, 2014.

[7] M. Xue, F. Kong, D. Weber, K. Thomas, Y. Wang, K. Brewster, K. Droegemeier, J. Weiss, D. Bright, M. Wandishin *et al.*, "Caps realtime storm-scale ensemble and high-resolution forecasts as part of the noaa hazardous weather testbed 2007 spring experiment," in *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred*, 2007.

[8] D. J. Gagne II, A. McGovern, J. Brotzge, M. Coniglio, J. Correia Jr, and M. Xue, "Day-ahead hail prediction integrating machine learning with storm-scale numerical weather models." in *AAAI*, 2015, pp. 3954–3960.

[9] K. A. Brewster, D. R. Stratman, and R. Hepper, "4-d visualization of storm-scale forecasts using vapor in the hazardous weather testbed spring forecasting experiment," *28th Conf. on Severe Local Storms, Portland, OR, Amer. Meteor. Soc., 15B.6.*

APPENDIX

Table III shows the average errors of a particular machine learning model which was better than other machine learning models that have been used to create aggregate model for each grid point. The errors were averaged over the locations where that particular machine learning model outperformed other machine learning models.

Table IV shows the average errors of a particular machine learning model which was better than other machine learning models that have been used to create aggregate model and also old conceptual models for each grid point. The errors were averaged over the locations where that particular machine learning model outperformed other machine learning and conceptual models.

Figure 6 plots the average precipitation prediction of conceptual model 1 for each real precipitation.

TABLE III
AVERAGE ERRORS WHEN A PARTICULAR MACHINE LEARNING MODEL WAS BETTER THAN OTHER MACHINE LEARNING MODELS

Machine Learning Models	RMSE
Linear Regression	442.437
K-Nearest Neighbors	316.123
Neural Network	83.108
Random Forest	78.164
Support Vector Machine	22.251
Polynomial Regression	253.283
Weighted Average	501.089

TABLE IV
AVERAGE ERRORS WHEN A PARTICULAR MACHINE LEARNING MODEL WAS BETTER THAN OTHER MACHINE LEARNING MODELS AND OLD MODELS AS WELL

Machine Learning Models	RMSE
Linear Regression	348.270
K-Nearest Neighbors	287.731
Neural Network	59.288
Random Forest	40.700
Support Vector Machine	2.717
Polynomial Regression	198.841
Weighted Average	477.335

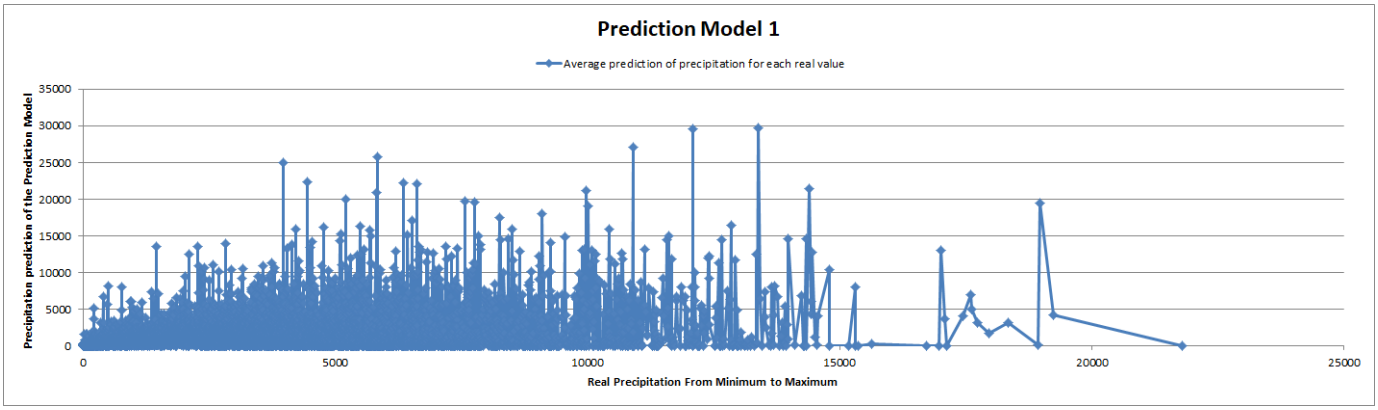


Fig. 6. Prediction Model 1 over prediction graph

Figure 7 plots the Root Mean Square Error of different prediction models for each grid point divided by the average precipitation.

Figure 8 shows the error of aggregate model that beats each individual old conceptual model in each grid point.

Figure 9 illustrates the error of aggregate model that beats each individual old conceptual model in each grid point divided by average precipitation.

Figure 10 plots the error of aggregate and conceptual models at each grid point.

Figure 11 shows the error of aggregate and conceptual models in each grid point divided by average precipitation.

Figure 12 illustrates the best model for each grid point which is linear in old models (checked if the best ML model is linear regression, k nearest neighbors or weighted average).

Figure 13 plots the best performing machine learning model comparing to other machine learning models at each grid point.

Figure 14 shows the best performing machine learning model comparing to other machine learning models as well as old prediction models at each grid point.

Figure 15 plots the error of the best performing machine learning model comparing to other machine learning models at each grid point.

Figure 16 illustrates the error of the best performing machine learning model comparing to other machine learning models as well as old prediction models at each grid point.

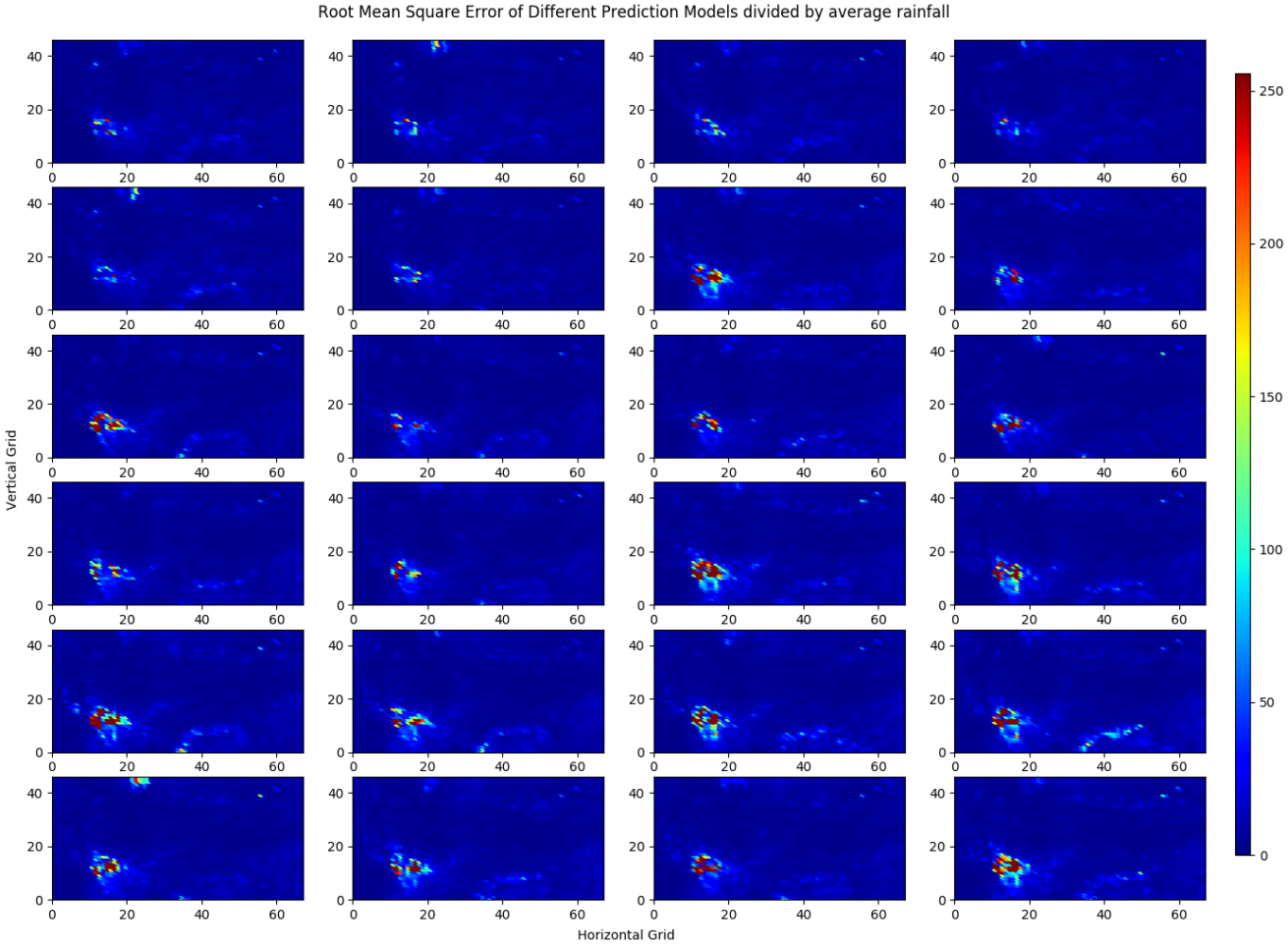


Fig. 7. Errors of Different Prediction Models divided by average

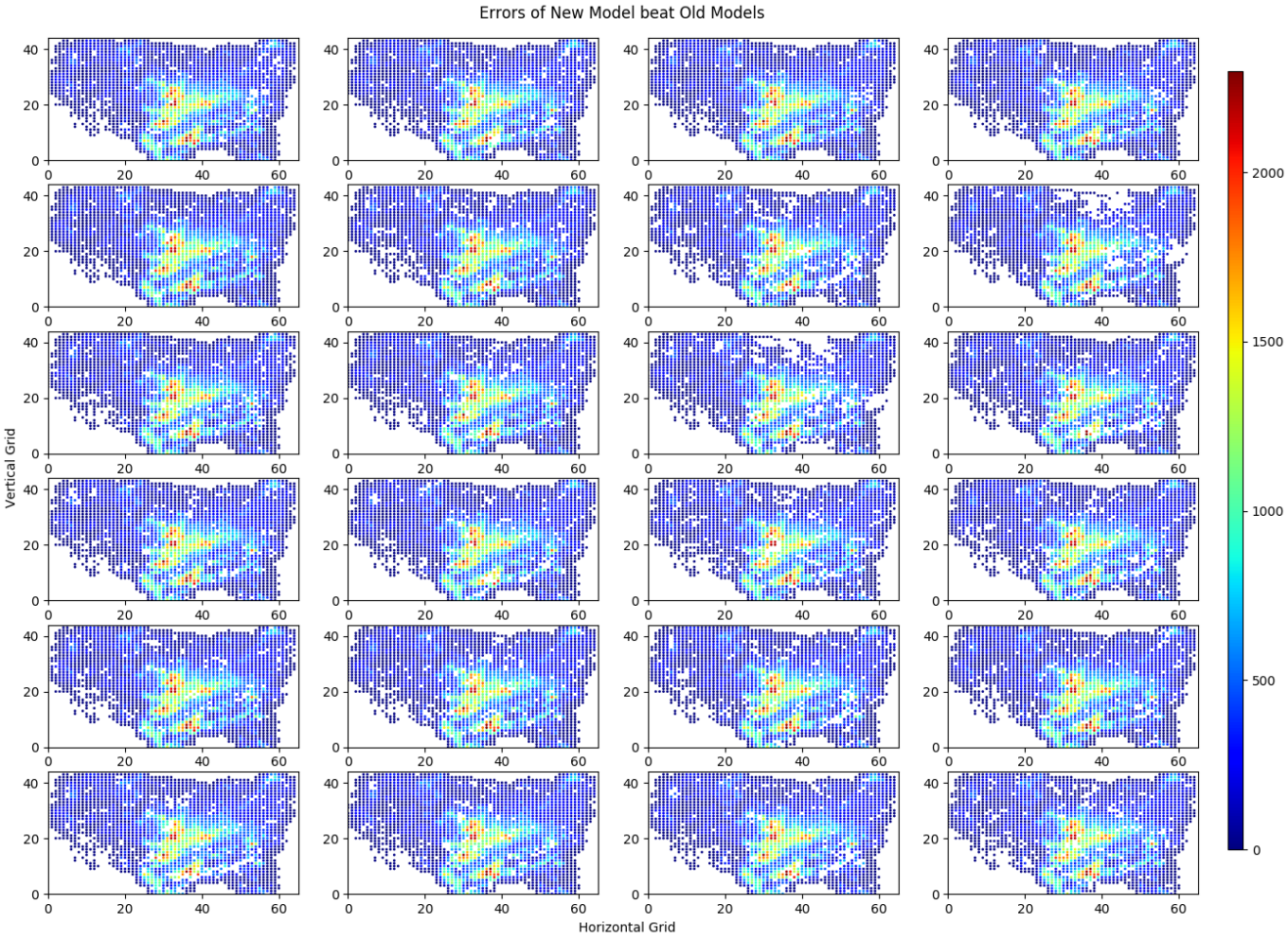


Fig. 8. Error of each new model beats old models

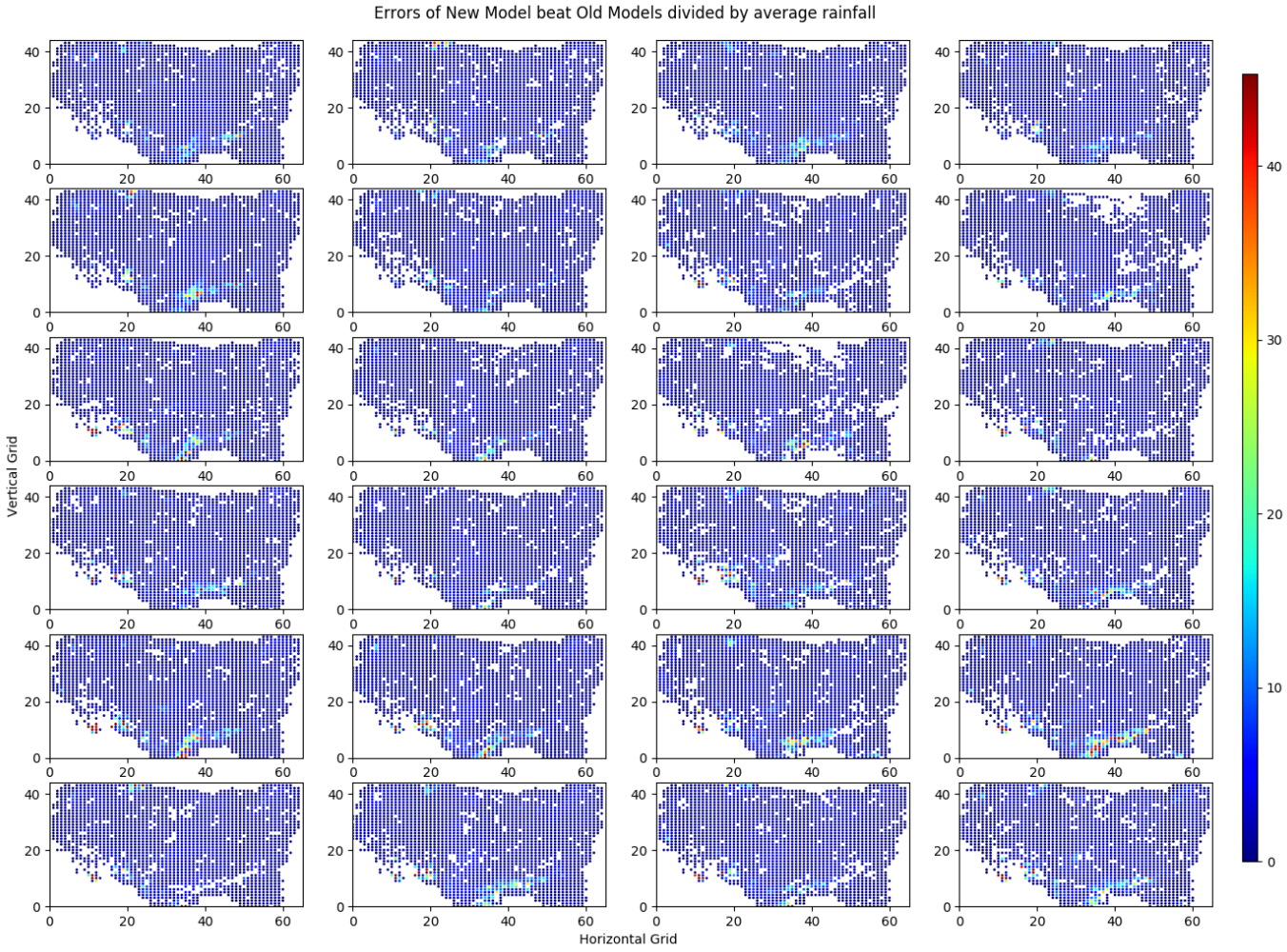


Fig. 9. Error of best aggregate model divided by average rainfall

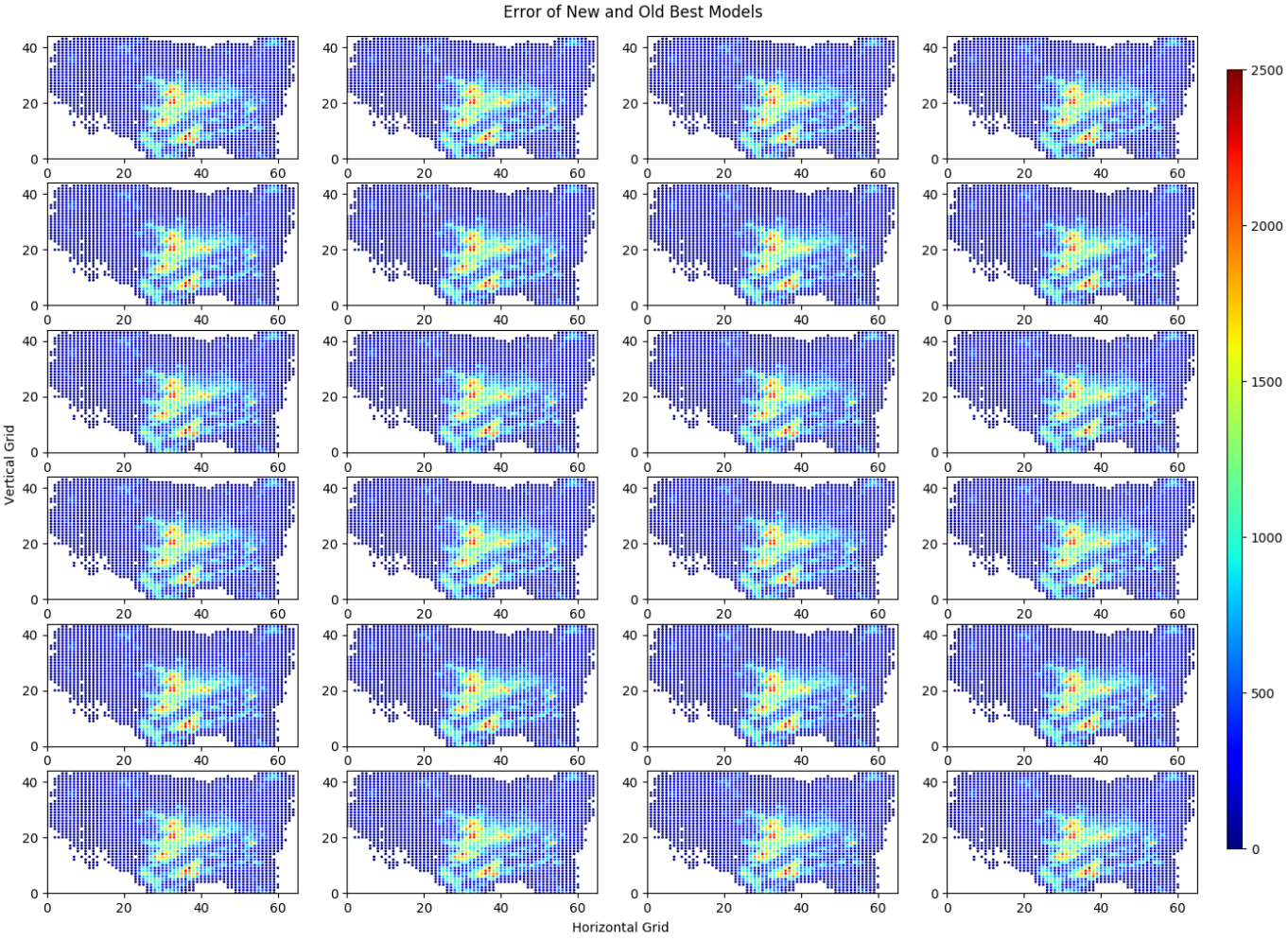


Fig. 10. Error of new and old models at each grid point

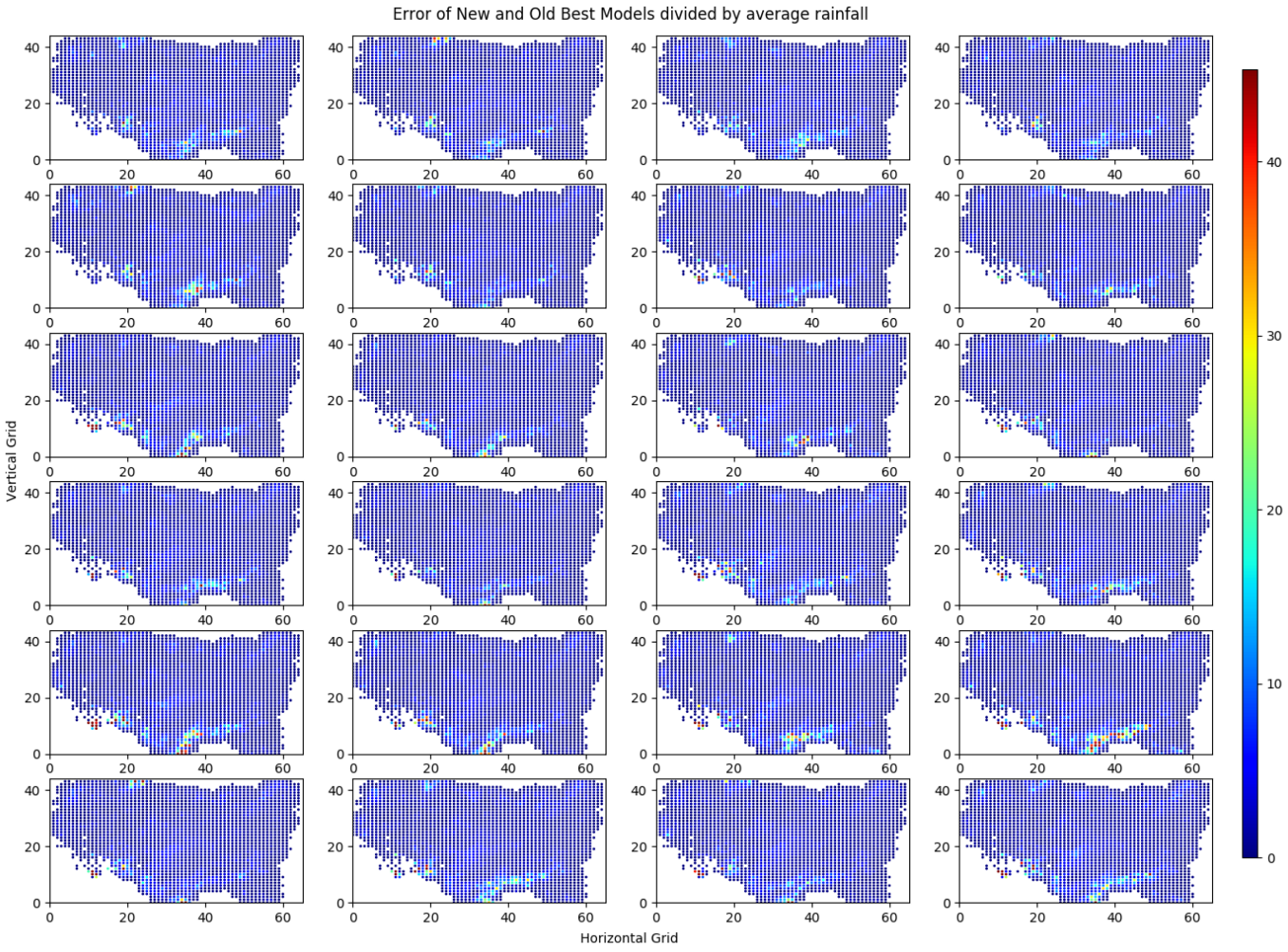


Fig. 11. Error of new and old models at each grid point divided by average

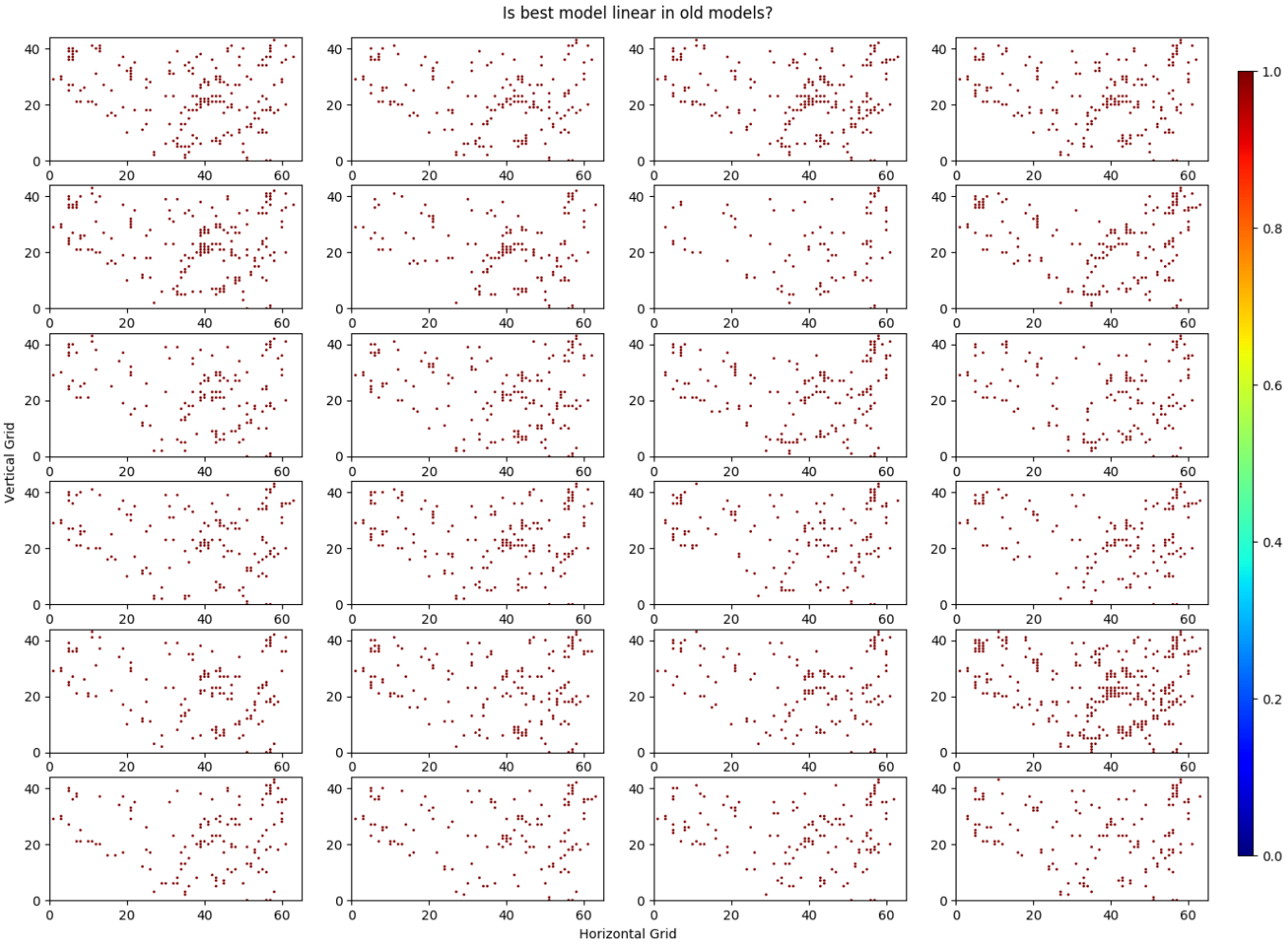


Fig. 12. The best machine learning models which are linear in old models at each grid point

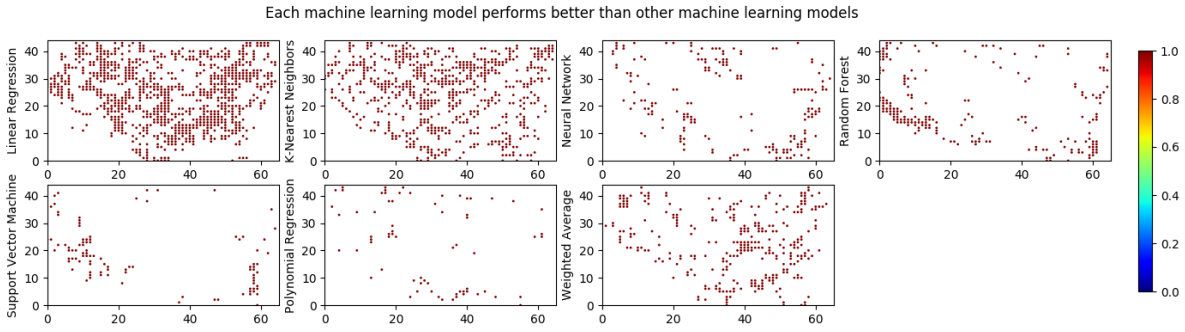


Fig. 13. Each machine learning model performs better than other machine learning model

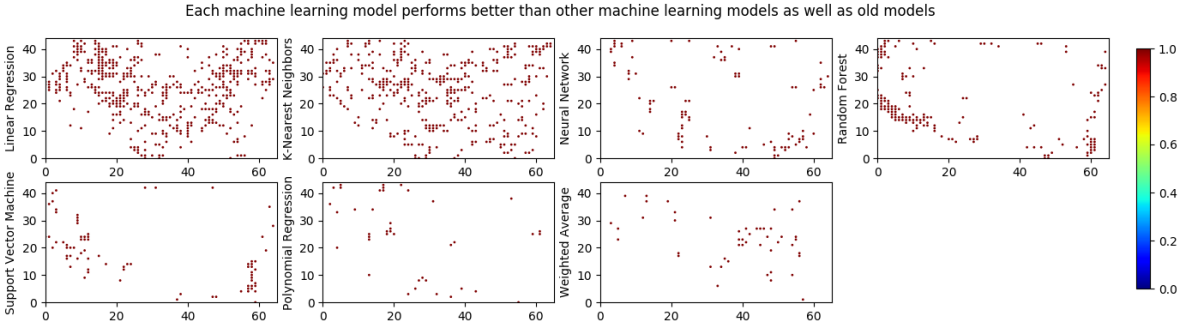


Fig. 14. Each machine learning model performs better than other machine learning model as well as old models

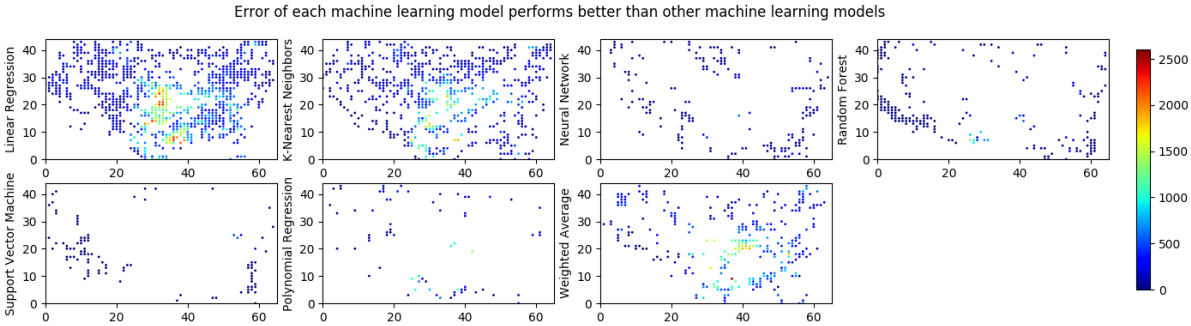


Fig. 15. Errors of each machine learning model performs better than other machine learning model

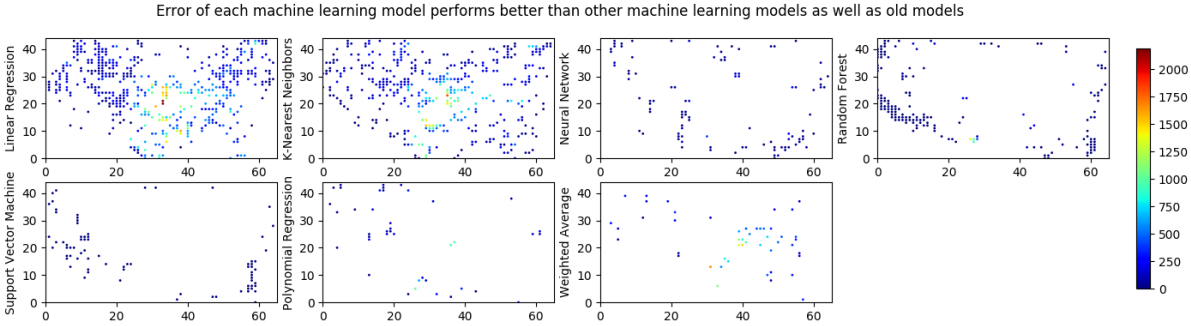


Fig. 16. Errors of each machine learning model performs better than other machine learning model as well as old models