Statistical and Machine Learning Methods Applied to the Prediction of Different Tropical Rainfall Types

Jiayi Wang[1], Raymond K. W. Wong[1], Mikyoung Jun[2], Courtney Schumacher[3], R. Saravanan[3], Chunmei Sun[2]

[1]Department of Statistics, Texas A&M University
[2]Department of Mathematics, University of Houston
[3]Department of Atmospheric Sciences, Texas A&M University

Key Points:

- The occurrence and intensity of deep convective, stratiform, and shallow convective rain can be predicted by each method to varying degrees.
- The generalized linear model (random forest) predicts rain (no-rain) occurrence best, the neural network produces best intensity statistics.
- None of the methods can predict the extreme tail (top 1%) of the rain rate distributions, although random forest comes closest.

Corresponding author: Jiayi Wang, jiayiwang@stat.tamu.edu

Abstract

We explore the use of three advanced statistical and machine learning methods (a generalized linear model, random forest, and neural network) to predict the occurrence and rain rate distribution of three tropical rain types (deep convective, stratiform, and shallow convective) observed by the radar onboard the GPM satellite over the West Pacific at three-hourly, 0.5-degree resolution. Temperature and moisture profiles from MERRA-2 were used as predictors. All three methods perform reasonably well at predicting the occurrence and rain rate distribution of each rain type. However, none of the methods obviously distinguish themselves from one another and each method still has issues with predicting rain too often and not fully capturing the high end of the rain rate distributions, both of which are common problems in climate models.

Plain Language Summary

Predicting rain from just large-scale environmental variables remains a challenging problem for climate models and it is unclear how well numerical methods of any kind can predict the true characteristics of rainfall without smaller (storm) scale information. The goal of this study was to explore the ability of multiple statistical and machine learning methods to predict rain occurrence and intensity over the tropical Pacific Ocean using satellite rain observations and large-scale environmental profiles of temperature and moisture. We also separated the rain into different types because of their varying kinematic and thermodynamic structures that might respond to the large-scale environment in different ways. Our expectation was that the machine learning methods (i.e., the neural network and random forest) would outperform the statistical model because of their more flexible structures, especially in predicting the highly skewed distribution of rain rates for each rain type. However, this was not the case for a standard neural network while the random forest produced a modest improvement over the statistical model. A possible moral of this story is that machine learning tools must be carefully assessed and are not necessarily applicable to solving all big data problems.

1 Introduction

Rainfall is fundamental to water resources, agriculture, and ecosystems and can cause massive damage in the form of too little or too much rain. However, rainfall can vary strongly in space and time making it hard to measure and even harder to predict. The rain rate distribution of most global climate models (GCMs) is far different than observed, with too much weak rain and not enough heavy rain (e.g., Dai, 2006; Stephens et al., 2010; Fiedler et al., 2020), which hinders predictions of extreme events. The goal of this study is to analyze the ability of advanced statistical and machine learning techniques to predict the occurrence and rain rate distribution of tropical rainfall using environmental temperature and humidity profiles as predictors. An eventual goal would be to determine if these techniques could be implemented in GCM predictions of short-term climate phenomena like El Niño, and perhaps even long-term climate change.

Most of the global rain falls in the tropics and warm season mid-latitudes and over half of this rain comes from large, organized rain systems (Nesbitt, Cifelli, & Rutledge, 2006; R. S. Schumacher & Rasmussen, 2020). These systems are much larger than the individual convective cells targeted by most conventional GCM convective parameterizations and contain elements of deep convection and stratiform rain (Houze, 1997; C. Schumacher & Houze, 2003a). An example of this kind of organized convective system is shown in Figure 1. Shallow convective rain is another type of rainfall that is ubiquitous over the tropical ocean and occurs regularly over some continental locations (C. Schumacher & Houze, 2003b; Funk, Schumacher, & Awaka, 2013).

As discussed by Mapes et al. (2006), these rain types form the building blocks of larger convective systems ranging from mesoscale convective systems (with scales on the order of 100 km and 12 h) to the Madden-Julian Oscillation (with scales on the order of 1000 km and many weeks), so predicting each of these rain types is important to studies of weather and climate. However, the ability of GCMs to simulate these building blocks and their interactions remains a challenge. For example, while deep convective rain is produced by GCMs via a convective parameterization, most GCMs produce shallow convection in their boundary layer parameterization, which is run separately from the convective parameterization. In addition, GCM convective parameterizations do not typically account for stratiform (or mesoscale) rain processes and the rest of a GCM's rainfall is produced explicitly at the grid scale as large-scale rain. It is important to note that large-scale rain in a GCM does not represent the observed stratiform building block discussed above.

Weather radar has the unique capability to view the 3-dimensional structure of precipitating storms (e.g., Figure 1), which can be used to determine the occurrence and evolution of the three tropical rainfall building blocks. Thus, this study utilizes spaceborne radar observations separated into deep convective, stratiform, and shallow convective rain to assess the predictive capability of advanced statistical and machine learning methods.

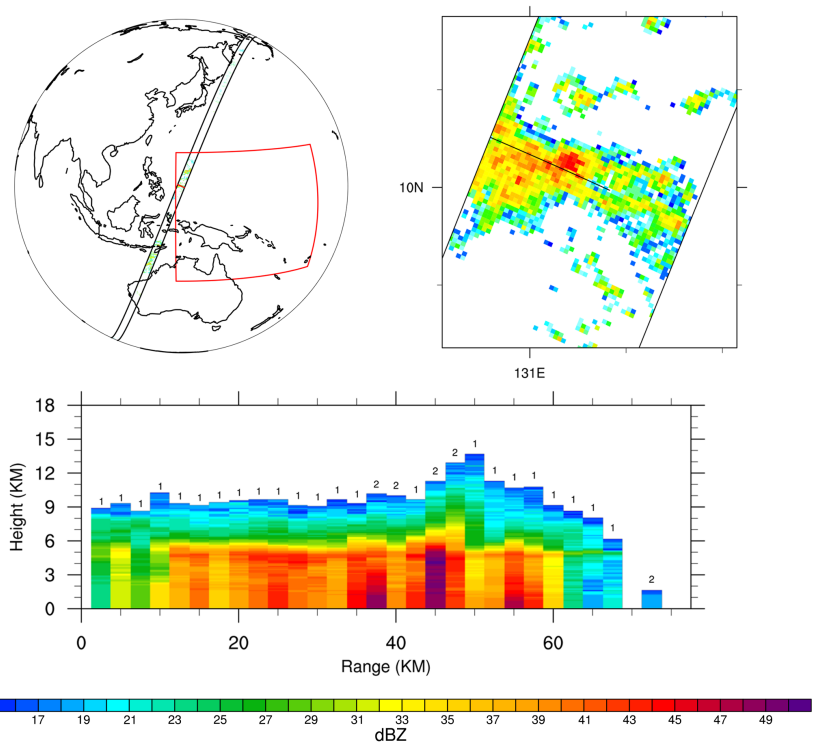

Figure 1. GPM DPR reflectivity observations at 01 UTC on 4 February 2017. The red box indicates the bounds of the study area over the West Pacific. The horizontal cross section is at 2 km AMSL and the vertical cross section is taken along the black line. Stratiform profiles are labeled as 1, convective profiles are labeled as 2. The far right cell in the vertical cross section would be considered shallow convection because its top is below the 0 degree Celsius level (typically about 5 km in the tropics).

There are currently a number of efforts to use data science to improve the representation of subgrid processes in climate models. Since there is often very limited amount of data available for unresolved processes, especially in situ measurements, many of these efforts apply machine learning to conventional model parameterizations or a large ensemble of higher resolution simulations (Brenowitz & Bretherton, 2018; O'Gorman & Dwyer, 2018; Rasp, Pritchard, & Gentine, 2018). Training on conventional parameterizations can improve computational efficiency, but does not address the physical deficiencies. The higher resolution simulations also have their own built-in assumptions about a different set of smaller scale unresolved processes. Yang et al. (2019) considered a data-centric approach, using a large satellite rainfall data set and reanalysis fields to show that a generalized linear model (GLM) can do well at predicting the occurrence of rain in the tropics, but it failed at capturing the tail of the rain rate distributions. This is mainly due to the restriction of parametric probability distributions used for rain rate. Although distributions such as Gamma, log-normal, or Weibull are commonly used for rain rate due to their shape of density curves with long tails (e.g., Yang et al. (2019) used a Gamma distribution), they are often not flexible enough to capture the heaviest rain rates. This study builds on Yang et al. (2019) by applying two machine learning techniques, i.e., a random forest (RF) and deep feedforward neural network (NN), to a similar data set to determine how well these methods compare to one another and the GLM in predicting rain occurrence and capturing the high rain rate end of the distribution for multiple rain types.

2 Statistical and Machine Learning Methods

2.1 Generalized Linear Model

GLMs (McCullagh & Nelder, 1989) are a popular class of statistical models used to predict a response variable whose mean is assumed to be some parametric function of covariates. It is a more general modeling framework than multiple linear regression in that response variables may not follow a Gaussian distribution. Furthermore, unlike multiple linear regression models, which often use the least squares method for model fitting, GLMs are fitted using a maximum likelihood estimation (MLE) method. The MLE method utilizes the distribution function of the response, thus giving generally better statistical properties of estimators than the least squares method. A GLM does not necessarily assume a direct linear relationship between the response and covariates, and often their nonlinear relationship is introduced by a *link* function. For instance, a common log-link function assumes that the log transformed mean of the response can be written as a linear combination of covariates. Widely used examples for distributions and link functions for GLMs include *logistic regression* (a Bernoulli distribution for the response and log link), *loglinear regression* (a Poisson distribution for the response and log link), and *Poisson regression* (a Poisson distribution for the response and log link).

In this work, we adopt the two-step modeling procedure used in Yang et al. (2019). Two separate GLMs, a logistic regression and a Gamma regression, are employed to deal with rain occurrence and rain amount, respectively. At a given time, let $p(\mathbf{s})$ denote the probability of rain at a grid point $\mathbf{s}$. Then the rain event is assumed to follow a Bernoulli distribution with

$$\log\Big\{\frac{p(\mathbf{s})}{1-p(\mathbf{s})}\Big\} = \beta_0 + \beta_1 z_1(\mathbf{s}) + \cdots + \beta_p z_p(\mathbf{s}), \tag{1}$$

where $z_i(\mathbf{s})$ denotes predictors (i.e. covariates) at the grid point $\mathbf{s}$. If $y(\mathbf{s})$ denotes the rain amount at $\mathbf{s}$, we assume that y follows a Gamma distribution with

$$\log[\mathrm{E}\{y(\mathbf{s})\}] = \eta_0 + \eta_1 z_1(\mathbf{s}) + \cdots + \eta_p z_p(\mathbf{s}). \tag{2}$$

For both models, parameters, including the coefficients β_i and η_i in (1) and (2), are estimated using the MLE method. We fit the GLM models using data aggregated over space

and time altogether, similar to Yang et al. (2019). Although models (1) and (2) do not have explicit temporal structure in them, the temporal structure of the covariates effectively account for that of the responses, and it did not seem necessary to add more temporal terms in (1) or (2).

Statistical inference on the estimated parameters, including the significance of coefficients, is made possible by using GLMs, and the estimated coefficients are readily interpretable. On the other hand, a possible drawback of the approach outlined above is the linearity assumption given in (1) and (2), as well as the distribution assumption on rain amount. In particular, the Gamma distribution may be too restrictive to account for some heavy rain events (Yang et al., 2019). Other commonly used distributions such as log-normal and Weibull distributions have similar problems, due to their particular parametric forms and restrictions. In view of the potentially restrictive nature of GLMs, we explore two popular machine learning methods, RF and artificial NNs, which operate under much weaker (i.e., non-linear) assumptions compared to GLMs. RF and NNs offer the most competitive predictive performances in many applications, and are now standard tools for machine learning.

2.2 Random Forest

Random forest (Breiman, 2001) is an ensemble learning method that makes predictions based on multiple decision trees. A random *forest* is built upon these many decision *trees*. A decision tree is a simple model that predicts the label associated with a sample by a series of splitting rules. An example decision tree is shown in Figure 2, where a tree is used to determine if a binary response Y is 1 or 0. The root node has a splitting condition: "$X_1 > 0$?" If the observation fulfills this condition, it will be passed to the next condition: "$X_2 < 10$?" Otherwise, the tree predicts $Y = 0$. The procedure is applied recursively until the tree reaches a prediction of Y. For the construction of a decision tree, we refer the readers to Breiman (2001). In the above example, the underlying goal is classification, where the response is categorical. Decision trees can also be modified to handle a regression problem, where the response is quantitative.

The core idea of ensemble methods like RF is to combine weak predictive models to achieve strong predictive performance. A RF is usually trained with two "random" ideas. The first is bagging – for each tree, the training set is formed by resampling from the original data set with replacement. The second is feature randomness – each tree in a RF is trained with a random subset of features. Bagging lowers variance while feature randomization reduces the dependence across trees. They are beneficial to ensemble learning. The prediction of the RF is obtained by a majority vote over the predictions of the individual trees.

Similar to the GLM analysis, a two-step modeling procedure was implemented for RF in our work. Namely, we trained an RF model on rain occurrence and another RF model on rain amount. For both models, we used the default setting of the "randomForest" function from the R package"randomForest", except that we restricted the number of decision trees to 100 when predicting rain amount in order to alleviate the computational burden. As opposed to GLM, RF is a nonparametric method and can produce a highly nonlinear regression function. On the other hand, it is significantly more difficult to interpret the results of the RF model, although RF provides a measure of variable importance. In practice, one might also examine individual classification trees within the random forest to understand the results.

2.3 Neural Network

In recent years, artificial NNs (especially those with deep architecture) have become one of the most prominent models for complicated functions. A NN is based on

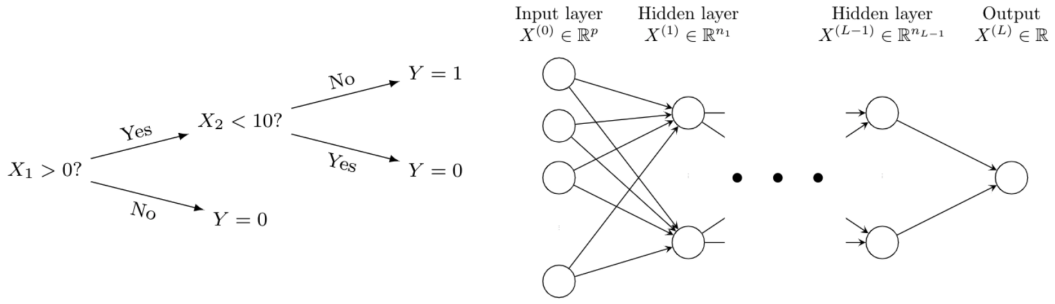

Figure 2. Illustrations for descision tree (left) and deep feedforward neural network (right).

a collection of connected nodes. Different ways to connect the nodes result in different NN architectures, such as fully connected (Hsu et al., 1990), sparsely connected (Ardakani et al., 2016), convolutional (Lo et al., 1995), and recurrent (Mikolov et al., 2010). Nodes are typically organized into layers, which can be classified as input, hidden and output. Networks with multiple hidden layers are said to have deep architectures, and are referred to as deep NNs. Deep architectures are commonly used nowadays, due to their strong empirical performance in many areas.

In our analysis, we adopt a deep feedforward NN in which consecutive layers are fully connected (Svozil et al., 1997; Schmidhuber, 2015) because it is one of the most standard forms of deep NN. Figure 2 depicts an example. We use $X^{(l)} \in \mathbb{R}^{n_l}$ to represent the nodes at layer l, where n_l is the number of nodes at layer l. Take $X^{(0)}$ as the input and $X^{(L)}$ as the output.The hidden and output layers are generated as follows. Let $x_k^{(l)}$ be the node k of layer l, where $l = 1, \ldots, L$ and $k = 1, \ldots, n_l$. Then

$$x_k^{(l)} = \sigma_k^{(l)}(b_k^{(l)} + \sum_{i=1}^{n_{l-1}} w_{i,k}^{(l)} x_k^{(l-1)}),$$

where $\sigma_k^{(l)}$ is the activation function, and $b_k^{(l)}$ and $w_{i,k}^{(l)}$ are parameters to be trained by the data. For simplicity, it is common to use the same activiations within the same layer: $\sigma^{(l)} := \sigma_k^{(l)}$, for $k = 1, ..., n_l$.

Similar to the previous two models (GLM and RF), we adopted the two-step approach for the NN analysis. More specifically, we trained one NN to perform the binary classification on rain occurrence and another NN using training samples with positive rain values only to predict the rain amount. We considered different number of layers for NN. More specifically, we considered $L = 2, 3, \ldots, 10$. Note that $n_0 = 80$ and $n_L = 1$ for all L since they are representing the input size and the output size. For any existing hidden layer, the number of nodes are set as follows: $n_1 = 40$, $n_2 = 20$, $n_3 = \cdots = n_{L-2} = 6$ and $n_{L-1} = 3$. For instance, for $L = 1$, there is only one hidden layer and so only n_1 is relevant. For $l = 1, \ldots, L-1$, the corresponding activation functions $\sigma_k^{(l)}$ were chosen as the rectified linear unit (ReLU) functions ($\sigma(x) = \max(0, x)$). The activation function for the output layer had to be chosen based on the response type, i.e., classification or regression. We used $\sigma^{(L)}(x) = 1/(1 + \exp(-x))$ for the classification, while we used the exponential function for the regression since the response is positive. As for the estimation of the NN, we adopted mean square error as the loss function and trained the network via the popular algorithm Adam (Kingma & Ba, 2014).

To prevent over-fitting, we also adopted the dropout procedure, which is a common regularization method for training deep neural networks (Baldi & Sadowski, 2013; Gal et al., 2017). In the dropout procedure, neurons are stochastically dropped out during the training at each layer. In our implementation, the dropout rate was set to be the

same at every layer and three possible values 0, 0.2, 0.5 were considered. Both the dropout rate and the number of layers, L, were regarded as the hyper-parameters and were chosen via a validation procedure — we randomly separated 20% of the training data as the validation set to select the best combination of dropout rate and number of layers.

3 Training and Test Data

We used two years of observations from the NASA Global Precipitation Measurement (GPM; Hou et al., 2014) dual-frequency precipitation radar (DPR) to calculate rain occurrence and rain rates, which were the predictands of the study. The full year of 2017 was used for training and the full year of 2018 was used for testing. The rain type classifications (i.e., deep convective, stratiform, and shallow convective; Funk et al., 2013) and associated rain rates were retrieved from 2ADPR v6 files. Figure 1 shows an example orbit from the GPM radar with all three rain types present. We regridded the DPR orbital rain observations, which are made at a 5-km footprint scale over a 245-km swath, to 0.5-degree horizontal resolution and 3-hourly temporal resolution. The predictors for the study were temperature and humidity fields at 40 pressure levels from the MERRA-2 reanalysis (Rienecker et al., 2011) for 2017 and 2018. The MERRA-2 data was regridded to a similar horizontal and temporal resolution as the DPR data and points were only analyzed if a DPR orbit occurred in a grid during the 3-hour period. We limited our domain to the tropical West Pacific ($130°\text{E} - 180°\text{E}$, $20°\text{S} - 20°\text{N}$; Figure 1), but found similar results in the tropical East Pacific (not shown). Overall, we had 569,596 training samples and 572,968 test samples.

The training and test data are generally similar to the observational data sets used in Yang et al. (2019). However, we used rain observations from the GPM DPR instead of the Tropical Rainfall Measuring Mission (TRMM) precipitation radar (PR) because of the DPR's higher sensitivity to weaker rain rates and thus better shallow convective rain retrievals (Hamada & Takayabu, 2016). We also used a slightly higher time resolution (3 hours vs 6 hours) to better isolate environment-rain relationships and we used all times of day instead of just 0-6 UTC to capture the full range of diurnal conditions. Finally, we only used temperature and humidity as predictors because they accounted for the majority of the predictive performance by the GLM in Yang et al. (2019), who also tested other environmental variables such as horizontal wind profiles and surface fluxes. We further utilized the full temperature and humidity profiles rather than just the first three empirical orthogonal functions so that the machine learning techniques had more flexibility in determining the vertical relationship of the predictors to the surface rain rate.

4 Prediction Results

4.1 Rain occurrence

When solving for occurrence, we treat grids with extremely small rain amounts as no-rain cases to avoid retrievals from the radar likely associated with clutter or noise. For each rain type, we selected a rain rate cutoff that accounts for less than 1% of the total rain amount in the training data. The cutoff values are 0.056, 0.0395, and 0.0087 mm/hr for deep convective, stratiform, and shallow convective rain, respectively. As will be illustrated in the next section, the three rain types produce different ranges of rain rate intensity, which is why separate cutoff values are needed for each rain type.

Rain does not occur often at the time and space scales being considered in this study (i.e., 3 hourly and 0.5 degrees), so there are many more no-rain cases than rain cases. To deal with this severely imbalanced classification problem, we created a "balanced" training data set by using a random under-sampling procedure. That is, we randomly sample the no-rain cases until we have the same number of no-rain and rain samples in

Table 1. [Table updated] The top four rows describe the performance of the occurrence predictions for each rain type by each method. The values in each column are the fraction of the total cases that fall into each prediction category and sum to one, while bold values are the highest correct predictions. The bottom two rows quanitify the accuracy of the the rain rate (mm/hr) prediction in terms of root mean square error (RMSE) and mean absolute error (MAE), with bold values representing the smallest errors among the three methods.

	Deep convective			Stratiform			Shallow convective		
	GLM	RF	NN	GLM	RF	NN	GLM	RF	NN
True Negative	0.485	**0.568**	0.536	0.474	**0.529**	0.502	0.325	**0.415**	0.323
False Negative	0.036	0.054	0.054	0.052	0.069	0.076	0.084	0.137	0.106
True Positive	**0.122**	0.103	0.103	**0.188**	0.171	0.164	**0.267**	0.214	0.245
False Positive	0.357	0.275	0.387	0.286	0.231	0.306	0.324	0.234	0.325
RMSE	0.758	0.975	**0.749**	0.624	0.730	**0.619**	0.095	0.105	**0.094**
MAE	0.405	0.504	**0.385**	0.295	0.367	**0.275**	**0.058**	0.062	0.059

our training data set. Note that we classify rain/no-rain cases for each rain type separately.

The top four rows of Table 1 show how well the three statistical and machine learning methods described in section 2 predict no-rain and rain cases for each rain type. The actual time the GPM radar observed each rain type over the West Pacific is indicated by adding the false negative and true positive values (i.e., about 16%, 24%, and 35% for deep convective, stratiform, and shallow convective rain, respectively). All three methods do a reasonable job at distinguishing truly raining cases, with GLM slightly outperforming the other two methods. However, all methods suffer from a relatively high false positive rate (i.e., predicting rain too often), which is a persistent problem in most climate models as well (Fiedler et al., 2020). While GLM had the best true positive predictions, it had the worst true negative predictions (i.e., predicting no rain when no rain is observed). RF had the best true negative prediction and NN fell between the two other techniques.

4.2 Rain rate distributions

We next apply the statistical and machine learning methods to predict the rain rate distribution of the three rain types. Figure 3 compares the prediction of each method to the "True" distribution observed by the GPM DPR. Note that the GPM-observed 99.9% rain rate varies by rain type with values of 14, 10, and 1.1 mm/hr for deep convective, stratiform and shallow convective rain, respectively. Even though shallow convective rain has the highest occurrence it has by far the smallest rain amounts over the 0.5-degree grid because it doesn't cover much of the grid and is composed of more lightly raining cells. Stratiform rain is also normally less intense than deep convective rain on a pixel-by-pixel basis but because it tends to cover more area than deep convective cells, stratiform rain amounts approach deep convective values over the 0.5-degree grid.

Figure 3 shows that all three methods (indicated by different green lines) tend to underestimate weaker rainrates (i.e., around the 50% quantile or first tick mark) in the deep convective and stratiform distributions, shifting to overestimations around the 90% quantile (or second tick mark). Between the 90 and 99% quantiles, there is a rapid drop off in prediction counts compared to the true distribution with NN and GLM showing the most rapid decrease. RF is the only technique to produce predictions past the 99.9% quantile for deep convective rain, the category associated with the most extreme rain amounts.

All methods do better predicting the shallow convective rain rate distribution with the drop-off in counts not occurring until after the 99% quantile.

To provide context on how the observed and predicted rain rate distributions in Figure 3 compare to standard GCM output, we obtained a year of data from the NCAR Community Atmospheric Model, version 5 (CAM5; Neale et al., 2013). We use model output for 2003 instead of 2018 because it was readily available. While there may be small year-to-year variations in the rain rate distributions over the West Pacific, we do not expect them to be large, especially since neither 2003 or 2018 experienced strong El Niño or La Niña events. The original rain rate data had a 25×25km grid resolution so we aggregated rain rates to a 0.5×0.5 degree grid resolution to match our analysis. Hourly total precipitation (PRECT) and convective (PRECC) precipitation rates were also aggregated into 3 hourly rain rates. We use PRECC to represent deep convective rain and the difference between PRECT and PRECC (PRECT-PRECC) to represent stratiform rain. GCMs do not typically calculate a separate shallow convective rain rate, but there are only small differences between the GPM convective deep rain rate distribution compared to when we combine the observed deep and shallow convective rain rate distributions (i.e., deep convective rain dominates the convective rain rate distribution in the West Pacific).

As seen in the top panel of Figure 3, CAM5 (indicated by the dashed yellow line) does not provide a good density estimation for deep convective rain (and is, in fact, close to the GLM and NN distributions). Recent work has shown that a stochastic version of the Zhang-McFarlane convective parameterization used in CAM5 can improve the deep convective rain rate distribution (e.g., Wang et al., 2021), but stochastic techniques are still not regularly implemented in standard GCM runs. CAM5 appears to characterize the stratiform rain distribution well (Figure 3, middle panel), although large-scale rain from GCMs and stratiform rain from radar are not considered to be be produced the same way (e.g., Dai, 2006), so caution must be taken in this comparison.

To further assess predicted rain amounts using GLM, RF, and NN, we calculated the following metrics to measure the performance of the techniques:

1. Root mean squared error (RMSE) $= \sqrt{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 / N}$ and
2. Mean absolute error (MAE) $= \sum_{i=1}^{N} |\hat{y}_i - y_i| / N$,

where y_i is the observed rain amount for the i-th sample, and $\hat{y}_i$ is the predicted rain amount for the i-th sample, for $i = 1, \ldots, N$. Here samples are aggregated over space and time, and thus there are a total of N samples for each rain type. Note that MAE is in general less sensitive to large values compared to RMSE. Table 1 shows that RF has the highest (and thus worst) RMSE and MAE among the three techniques for each rain type. NN usually provides the smallest errors among the three methods, and GLM usually performs only slightly worse than NN.

5 Conclusions

There is strong motivation to use "big data" to parameterize unresolved processes in GCMs, such as rainfall production. While training and testing data can come from higher resolution models, we chose to use a multi-year data set of rain observations from satellite radar along with temperature and humidity fields derived from a model constrained by observations (i.e., reanalysis). There are also a number of advanced statistical and machine learning techniques with which to analyze the available data. We chose a representative set that ranged in ease of implementation and interpretability: a generalized linear model, random forest, and neural network.

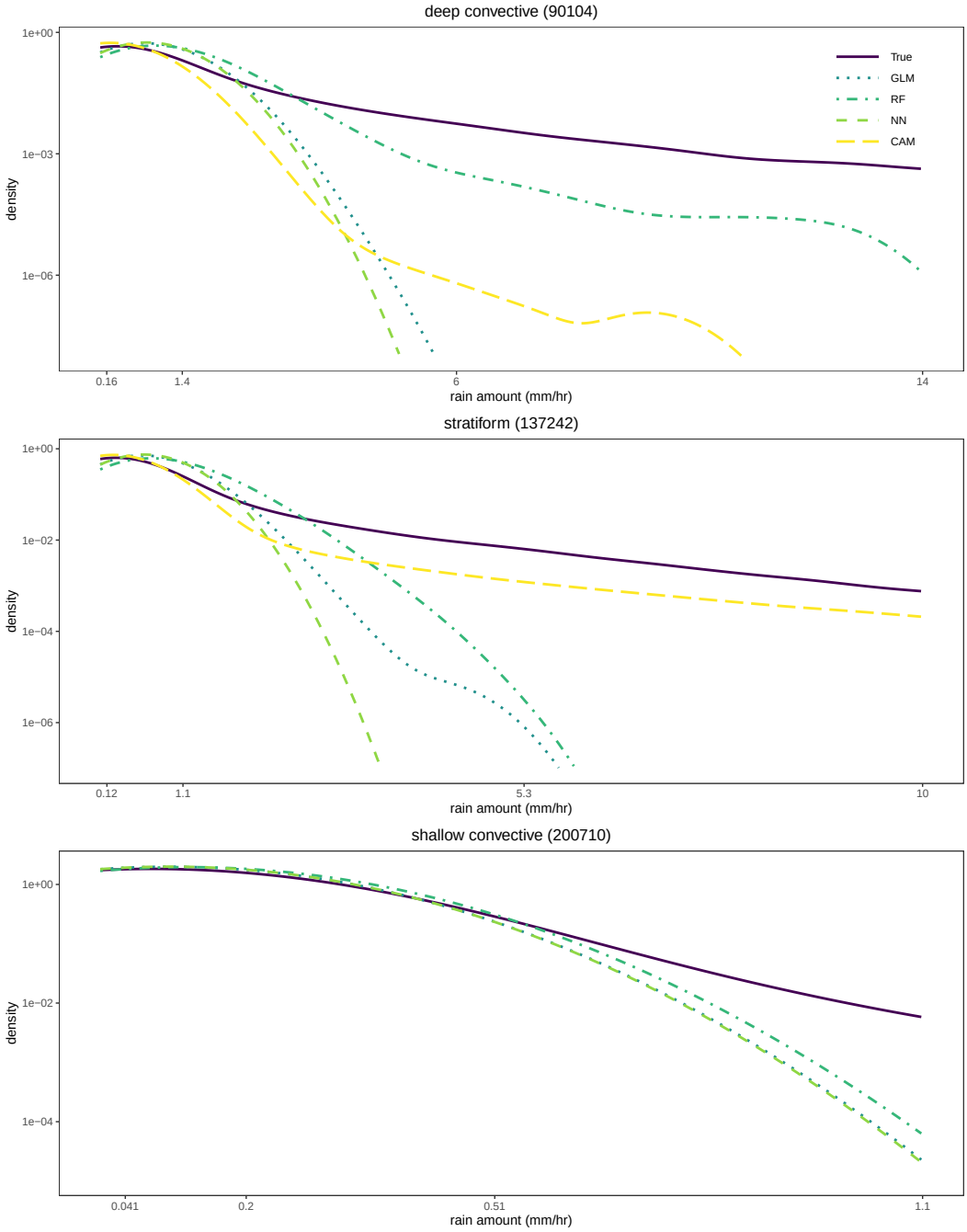

Figure 3. [Figure updated] GPM-observed and model-predicted rain rate distributions for deep convective, stratiform, and shallow convective rain in the base-10 log scale. Values in parentheses are the total cases in the testing data that rain. Values on the x-axis for the three plots are the 50, 90, 99, and 99.9% quantiles of the rain rate distribution, respectively.

All three methods performed reasonably well in predicting the occurrence of each of the three tropical building block rain types: deep convective, stratiform, and shallow convective. However, each method still predicted rain too often. Due to the high complexity of the model structure, regularization is usually needed for NN. With the dropout regularization, NN performed similarly to GLM in predicting the rain rate distributions of each rain type, while RF was more flexible in modeling the true response. However, the very highest rain rates were still underpredicted by all methods. Future work will assess the ability of each method to capture the temporal and spatial distribution of observed tropical rainfall, with the ultimate goal of implementing the best overall technique in a GCM to improve the representation of convection.

Acknowledgments

RW acknowledges support by NSF grants DMS-1711952 and DMS-1806063, and NASA grant 80NSSC19K0656. MJ acknowledges support by NSF grants DMS-1925119 and DMS-2105847, and NIH grant P42ES027704. CS acknowledges support by NASA grant 80NSSC19K0734. RS acknowledges support from DOE grant DE-SC0020072. The authors also acknowledge T3 grant (#246502) from Texas A&M University. The original data files for GPM and MERRA-2 can be acquired from the Goddard Earth Science Data Information Services Center (GES DISC) (https://disc.gsfc.nasa.gov/). Aaron Funk processed the GPM DPR and MERRA-2 data onto coincident temporal and spatial grids. Yangyang Xu provided the CAM5 data used for rain rate comparison.

References

Ardakani, A., Condo, C., & Gross, W. J. (2016). Sparsely-Connected Neural Networks: Towards Efficient VLSI Implementation of Deep Neural Networks. *arXiv preprint arXiv:1611.01427*.

Baldi, P., & Sadowski, P. J. (2013). Understanding Dropout. *Advances in neural information processing systems*, *26*, 2814–2822.

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5–32.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic Validation of A Neural Network Unified Physics Parameterization. *Geophys. Res. Lett.*, *45*, 6289–6298.

Dai, A. (2006). Precipitation Characteristics in Eighteen Coupled Climate Models. *Journal of climate*, *19*(18), 4605–4630.

Fiedler, S., Crueger, T., D'Agostino, R., Peters, K., Becker, T., Leutwyler, D., ... others (2020). Simulated Tropical Precipitation Assessed across Three Major Phases of the Coupled Model Intercomparison Project (CMIP). *Monthly Weather Review*, *148*(9), 3653–3680.

Funk, A., Schumacher, C., & Awaka, J. (2013). Analysis of Rain Classifications Over the Tropics by Version 7 of the TRMM PR 2A23 Algorithm. *Journal of the Meteorological Society of Japan. Ser. II*, *91*(3), 257–272.

Gal, Y., Hron, J., & Kendall, A. (2017). Concrete Dropout. *arXiv preprint arXiv:1705.07832*.

Hamada, A., & Takayabu, Y. N. (2016). Improvements in Detection of Light Precipitation with the Global Precipitation Measurement Dual-Frequency Precipitation Radar (GPM DPR). *Journal of atmospheric and oceanic technology*, *33*(4), 653–667.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., ... Iguchi, T. (2014). The Global Precipitation Measurement Mission. *Bulletin of the American Meteorological Society*, *95*(5), 701–722.

Houze, R. A., Jr. (1997). Stratiform Precipitation in Regions of Convection: A Meteorological Paradox? *Bulletin of the American Meteorological Society*, *78*(10), 2179–2196.

Hsu, K.-Y., Li, H.-Y., & Psaltis, D. (1990). Holographic Implementation of A Fully Connected Neural Network. *Proceedings of the IEEE*, *78*(10), 1637–1645.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lo, S.-C., Lou, S.-L., Lin, J.-S., Freedman, M. T., Chien, M. V., & Mun, S. K. (1995). Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection. *IEEE transactions on medical imaging*, *14*(4), 711–718.

Mapes, B., Tulich, S., Lin, J., & Zuidema, P. (2006). The Mesoscale Convection Life Cycle: Building Block or Prototype for Large-scale Tropical Waves? *Dynamics of atmospheres and oceans*, *42*(1-4), 3–29.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). Chapman & Hall/CRC, Boca Raton, Florida.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent Neural Network Based Language Model. In *Eleventh annual conference of the international speech communication association.*

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., & Zhang, M. (2013). The Mean Climate of the Community Atmosphere Model (CAM4) in Forced SST and Fully Coupled Experiments. *Journal of Climate*, *26*(14), 5150–5168.

Nesbitt, S. W., Cifelli, R., & Rutledge, S. A. (2006). Storm Morphology and Rainfall Characteristics of TRMM Precipitation Features. *Monthly Weather Review*, *134*(10), 2702–2721.

O'Gorman, P. A., & Dwyer, J. G. (2018). Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563.

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep Learning to Represent Subgrid Processes in Climate Models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., ... others (2011). MERRA: NASA's Modern-era Retrospective Analysis for Research and Applications. *Journal of climate*, *24*(14), 3624–3648.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural networks*, *61*, 85–117.

Schumacher, C., & Houze, R. A., Jr. (2003a). Stratiform Rain in the Tropics as Seen by the TRMM Precipitation Radar. *Journal of Climate*, *16*(11), 1739–1756.

Schumacher, C., & Houze, R. A., Jr. (2003b). The TRMM Precipitation Radar's View of Shallow, Isolated Rain. *Journal of Applied Meteorology*, *42*(10), 1519–1524.

Schumacher, R. S., & Rasmussen, K. L. (2020). The Formation, Character and Changing Nature of Mesoscale Convective Systems. *Nature Reviews Earth & Environment*, 1–15.

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J.-C., Bodas-Salcedo, A., ... Haynes, J. (2010). Dreary State of Precipitation in Global Models. *Journal of Geophysical Research: Atmospheres*, *115*(D24).

Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to Multi-layer Feed-forward Neural Networks. *Chemometrics and intelligent laboratory systems*, *39*(1), 43–62.

Wang, Y., Zhang, G. J., Xie, S., Lin, W., Craig, G. C., Tang, Q., & Ma, H.-Y. (2021). Effects of Coupling a Stochastic Convective Parameterization With the Zhang–McFarlane Scheme on Precipitation Simulation in the DOE E3SMv1.0 Atmosphere Model. *Geoscientific Model Development*, *14*(3), 1575–1593.

Yang, J., Jun, M., Schumacher, C., & Saravanan, R. (2019). Predictive Statistical Representations of Observed and Simulated Rainfall Using Generalized Linear

Models. *Journal of Climate*, *32*(11), 3409–3427.