Multi-Layer Networks for Ensemble Precipitation Forecasts Postprocessing

Fengyang Xu, Guanbin Li, Yunfei Du*, Zhiguang Chen, Yutong Lu
School of Computer Science and Engineering, Sun Yat-sen University
xufy9@mail2.sysu.edu.cn liguanbin@mail.sysu.edu.cn {yunfei.du, zhiguang.chen, yutong.lu}@nscc-gz.cn

Abstract

The postprocessing method of ensemble forecasts is usually used to find a more precise estimate of future precipitation, because dynamic meteorology models have limitations in fitting fine-grained atmospheric processes and precipitation is driven more often by smaller-scale processes, while ensemble forecasts can hit this precipitation at times. However, the pattern of these hits cannot be easily summarized. The existing objective postprocessing methods tend to extend the rain area or false alarm the precipitation intensity categories. In this work, we introduce a multi-layer structure to simultaneously reduce the bias in forecast ensembles output by meteorology models and merge them to a quality deterministic (single-valued) forecast using cross-grid information, which differs quite dramatically from the previous statistical postprocessing method. The multi-layer network is designed to model the spatial distribution of future precipitation of different intensity categories (IC-MLNet). We provide a comparison of IC-MLNet to simple average as well as another two state-of-the-art ensemble quantitative precipitation forecasts (QPFs) postprocessing approaches over both single-model and multi-model ensemble forecasts datasets from TIGGE. The experimental results indicate that our model achieves superior performance over the compared baselines in precipitation amount prediction as well as precipitation intensities categories prediction.

Introduction

Weather forecasting is usually solved by numerical weather prediction (NWP) models. Among all the atmospheric quantities predicted by the NWP model, precipitation is one of the most difficult to predict accurately. Errors in QPFs may arise due to errors in the observations and the forecasting model itself. Detailed knowledge of atmospheric moisture and vertical motion fields is essential for predicting the location and amount of precipitation, but these are difficult to observe accurately. The NWP models cannot adequately represent the cloud dynamics and microphysical processes involved in precipitation generation(Ebert 2001).

To compensate for shortcomings in model physics, weather centers have run a variety of carefully designed NWP models to generate forecasts of future states of the atmosphere(Flowerdew 2012), knowing that different forecasts are likely to pick up different nuances in the predicted weather, which formed ensemble forecasts. A *single-model ensemble forecast* is a collection of forecasts from an NWP model using slightly different initial conditions and model variants(Gneiting 2014), which is based on the implicit assumption that errors result primarily from uncertainties in the initial conditions. A *multi-model ensemble forecast* is a collection of outputs from different existing NWP models, which avoids the problem of systematic bias that occurs when a single model is used and in addition to having negligible cost. The average of either kind of ensemble usually has better skill, consistency, quality, and economic value than other single-valued NWP forecasts of similar grid resolution(Richardson 2000; Wandishin et al. 2001; Zhu et al. 2002; Kalnay 2003), except for rare events(Hamill et al. 2000).

However, the increasing number of ensemble members still cannot reduce errors caused by insufficient detailed atmospheric observations. Therefore, the calibration of NWP ensemble forecasts, as part of the precipitation postprocessing, has become a necessary and crucial step for the daily operational runs at numerical weather prediction centers. In fact, forecasters with rich practical experience at the bench routinely use multiple models for guidance, and this calibration is simply a formalization of this process.

In this paper, we use the real observation data of historical time to calibrate the ensemble forecasts (see Figure 1) and merge them to a "best estimate". A significant challenge in our study is that precipitation forms aloft, which is strongly

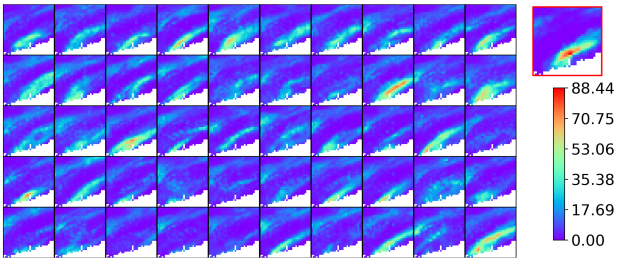


Figure 1: An ensemble forecast with 50 members (input) and the corresponding observation (in red block, ground truth)

*The author is the corresponding author.

affected by atmospheric vertical motions and depends on local variables such as wind at various elevations, temperature, humidity, and atmospheric pressure, which cannot be captured by atmospheric models and reflected in NWP forecasts. Because of this complexity, precipitation events of different ranges and different intensities have completely different causes of formation: large areas of precipitation rely on large-scale air masses, while thunderstorms occur locally in rising air surrounded by falling air. Therefore, it is difficult to simultaneously ensure the accuracy of the prediction of the light precipitation area and that of the extreme precipitation amount.

Inspired by the spatial fractal feature of precipitation, we propose a novel deep neural network, called IC-MLNet (Intensity Categories Multi-Layer Networks) for producing a deterministic precipitation forecast from NWP ensemble precipitation forecasts, based on the following two types of data: 1) the precipitation observation data collected by automatic weather stations; 2) the ensemble forecasts data from TIGGE (The International Grand Global Ensemble) dataset. In view of the complex triggers of precipitation, we provide a multi-layer network to reduce bias for light precipitation, moderate precipitation, heavy precipitation and violent precipitation separately, then merge the calibrated results. In addition, as mentioned above, the diversity of ensemble can compensate for the NWP errors to a certain extent, we try to learn member context features of ensemble forecasts. The non-local block instead of the traditional convolution was utilized to capture the long-range dependencies in each calibration layer.

The contribution of this work is threefold:

- We present IC-MLNet, an multi-layer neural network for ensemble precipitation forecasts postprocessing, which simultaneously reduce the bias in ensemble forecasts and merge them to a deterministic forecast. IC-MLNet outperforms the widely used ensemble average as well as a suite of state-of-the-art statistical ensemble postprocessing methods, in terms of precipitation's amounts, distribution and intensity.
- The ability of IC-MLNet to outperform the previous methods suggests the ability to leverage spatial information and member context information in terms of generating the most likely deterministic forecast from ensemble forecasts, while the improved ability beyond compared single-layer structure points to the possibility of learning the features of precipitation with different causes.
- IC-MLNet provides forecasters a scalable objective method of choice to generate a deterministic forecast based on ensemble forecasts (a single-value forecast regarded as a single member ensemble forecast) from different climate modeling groups across the world, run with different initial conditions.

Related Work

A common example of calibrating and integrating ensemble forecasts is Weighted Ensemble Average, where each ensemble member is weighted inversely to its past forecast error. (Zhi, Zhou, and Xu 2011) applied multi-model superensemble (Krishnamurti et al. 2000; Krishnamurti, Gnanaseelan, and Chakraborty 2007) approach to ensemble average, proposed Weighted bias-removed Ensemble mean (WEM), and verified its performance using TIGGE datasets. Although WEM tries to explore the relationship between observations and ensemble forecasts in historical data, abnormal results often occur (e.g., the precipitation in one point is much larger than the surrounding area) as it is a linear non-parametric method based on a single point.

The postprocessing methods run in operational centers usually take the experience of forecasters into account. ENSemble pseudo-Bias-Corrected QPF(ENSBC) (Novak et al. 2014) permits forecasters adjust calculated weights, and Best Percentile proposed by (Dai et al. 2016) is an improved approach of Probability Matching (PM) (Ebert 2001) which use the percentile values chosen by forecasters based on precipitation intensities rather than median to replace the sorted ensemble average value. These methods are flexible but unstable, because the skill of forecasters varies and only a limited number of members can be referred. In addition, they tend to choose the parameters that will produce more severe precipitation, which increase the false alarm rate of results. Also, these approaches are based on the single grid, ignoring the spatial local differences of precipitation.

Methods

Problem Definition

Ensemble precipitation calibration and integration aims at generating the most likely deterministic precipitation forecast from the corresponding ensemble forecasts. Generally, the deterministic forecast $Y \in \mathbb{R}^{h \times w}$ is modeled as the output of the following transforming:

$$Y = \mathcal{F}(\mathbf{X}; \Theta) \tag{1}$$

where $\mathbf{X} = \{x_1, \cdots, x_M\}$ is the ensemble forecasts with M NWP models and N_i forecasts per model, i.e. $x_i \in \mathbb{R}^{h \times w \times N_i}$. $\mathcal{F}$ is the postprocessing model, here is IC-MLNet, and Θ denotes the parameters of $\mathcal{F}$. The precipitation observation is used as the ground truth.

Basic Network Architecture

The proposed IC-MLNet (as shown in Figure 2) is a multi-layer structure: precipitation with different intensity is processed respectively, and each layer has a similar structure for extracting features of different intensity categories.

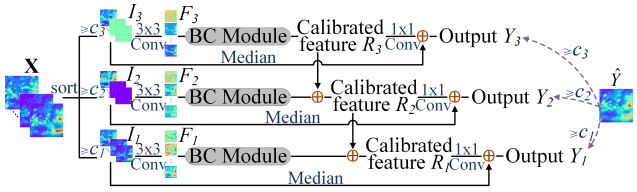


Figure 2: IC-MLNet Architecture ($L = 3$). BC module is bias correction module. $\oplus$ denotes element-wise sum. The purple dashed lines represent supervised pairs.

First, the input ensemble is stacked and sorted in the member dimension before feeding into the network. Sorting among members is a routine operation of previous nonparametric methods, in order to eliminate the interference of the member order on the results, and it does not reduce precipitation events. We do this also for the sake of scalability, and do not want the network to learn the dependencies between members for a certain dataset.

$$X = sort(stack(\mathbf{X})) \tag{2}$$

here $sort$ and $stack$ represent the sorting and stack operation. $X \in \mathbb{R}^{h\times w\times n}$ and $n = \sum_{i=1}^{M} N_i$ is the total number of input forecasts.

The basic network consists of L ($L = 1, 2, ...$) layers. Each layer has a bias correction module to generate the precipitation scenario residual of corresponding intensity, and then we fuse these fractal features and output the final deterministic forecast. This procedure can be expressed as:

$$F_l = f_{ext}(I_l),\ I_l = X \geq c_l \tag{3}$$

$$R_l = \mathcal{M}_l(F_l) + \alpha R_{l+1},\ l < L \qquad R_L = \mathcal{M}_L(F_L) \tag{4}$$

$$Y_l = \beta f_{fuse}(R_l) + median(I_l) \tag{5}$$

where $\geq$ represents the element-wise greater than, and c_l represents the intensity threshold of precipitation category. We use general thresholds of "light rain", "moderate rain", "heavy rain", and "violent rain" in operational centers here, that is: $c_1 = 0.1mm \cdot day^{-1}$, $c_2 = 10.1mm \cdot day^{-1}$, $c_3 = 25.1mm \cdot day^{-1}$ and $c_4 = 50.1mm \cdot day^{-1}$. $\mathcal{M}_l$ denotes the bias corrected module in l-th layer, $I_l \in \mathbb{R}^{h\times w\times n}$, $F_l \in \mathbb{R}^{h\times w\times c}$ and $Y_l \in \mathbb{R}^{h\times w}$ are the layer input, extracted feature and output. $R_l \in \mathbb{R}^{h\times w\times c}$ denotes the forecast residual of the layer, which is added with median of I_l, and the optimal weights α, β are automatically learned during training. f_{ext}, f_{fuse} denote the feature extraction function and fusion function, here (see Figure 2) are a 3×3 convolutional layer and a 1×1 convolutional layer. Therefore, our final output of IC-MLNet is

$$Y = \mathcal{F}(\mathbf{X}; \Theta) = Y_1 \tag{6}$$

Given a training set $\left\{\mathbf{X}^{(i)}, \hat{Y}^{(i)}\right\}_{i=1}^{K}$, where K is the number of training samples (batch size) and $\hat{Y}^{(i)}$is the ground truth observation of the ensemble forecasts $X^{(i)}$, An MSE loss function is used where the optimization objective is defined as:

$$\arg\min_{\Theta}(\mathcal{L}(\Theta)) \tag{7}$$

$$\mathcal{L}(\Theta) = \frac{1}{K}\sum_{i=1}^{K}\left\|\hat{Y}^{(i)} - \mathcal{F}(X^{(i)}; \Theta)\right\|_2^2 \tag{8}$$

here we are required to learn the parameter set Θ of basic end-to-end IC-MLNet.

Bias Correction Module

We now present the details of our bias correction module.

Bias correction module used to model a non-linear function calibrate the error between ensemble and ground truth

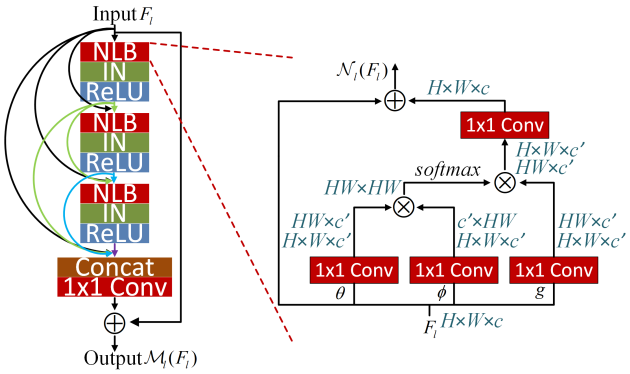


Figure 3: Details of bias correction module (left) and non-local block (NLB) $\mathcal{N}_l$. F_l is input feature, $\otimes$ and $\oplus$ denotes matrix multiplication and element-wise sum, respectively. We set $c' < c$ to reduce computation, and gray fonts represent the changing shapes of the features

using long-range spatial dependence. Here, we use the residual dense structure(Zhang et al. 2018) including non-local blocks, instance normalization (IN) and ReLU (see Figure 3 (left)). Non-local block is introduced in Non-Local Network (Wang et al. 2018) and shows powerful learning ability for video classification and image recognition. Non-local block can capture long-range and channel-wise dependencies directly by computing interactions between any two positions. For this multi-layer structure, the feature context information is an important reference for distinguishing different precipitation intensities, non-local block can help each layer generate features that better matches their intensity categories.

As shown in Figure 3, all convolution layers in the original residual dense structure are replaced with non-local blocks, and for the great differences among samples, we use instance normalization. In the non-local block, we simply use the 1×1 convolution in space dimension as unary function, the Embedded Gaussian function as the pairwise function, and the $softmax$ function as normalization factor.

Multi-supervised IC-MLNet

To further explore the features at different precipitation intensities, inspired by (Tai et al. 2017), we supervise outputs from all layers during training (Figure 2). The loss function of our multi-supervised IC-MLNet can be formulated as,

$$\hat{Y}_l = \hat{Y} \geq c_l \tag{9}$$

$$\begin{aligned}\mathcal{L}(\Theta) = &\frac{w}{K}\sum_{i=1}^{K}\left\|\hat{Y}_1^{(i)} - Y_1^{(i)}\right\|_2^2 \\ &+ \frac{1-w}{K}\sum_{i=1}^{K}\left\|\hat{Y}_l^{(i)} - \sum_{l=2}^{L} Y_l^{(i)}\right\|_2^2\end{aligned} \tag{10}$$

where w denotes the loss weight, which is a hyperparameter determined by the distribution of precipitation values. As

shown in Figure 4, light precipitation events in our datasets account for 0.77 of all precipitation events, i.e. $w = 0.77$.

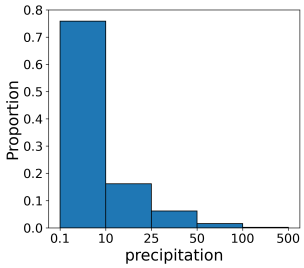


Figure 4: Distribution of daily precipitation amounts in single-model dataset

Experiments

This section describes the experiments performed to demonstrate the effectiveness of IC-MLNet when applied to a single-model and a multi-model ensemble forecasts dataset.

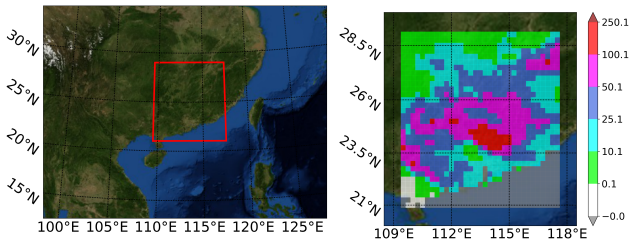
Datasets

Ensemble Forecasts Data The ensemble data we used is from TIGGE dataset[1]. Our single-model dataset (**Smod**) contains 50 members, which is an ensemble forecast from the ECMWF Center, while multi-model dataset (**Mmod**) consists of 4 deterministic forecasts from UKMO, NCEP, JMA and ECMWF. Both datasets have $0.25° \times 0.25°$ grid data in the area $[21.0°\text{N} \sim 29.0°\text{N}, 109.5°\text{E} \sim 117.5°\text{E}]$ (as shown in Figure 5(a)). 366-hour forecasts are released by TIGGE at UTC0000 and UTC1200 every day, but only the 24-h total precipitation in $6 \sim 30$, $12 \sim 36$, $18 \sim 42$ and $24 \sim 48$ forecast hours are used.

Observation Data Observation data is collected from 7247 automatic stations[2] in $[21.0°\text{N} \sim 29.0°\text{N}, 109.5°\text{E} \sim 117.5°\text{E}]$, and spans January 2013–December 2016 (except 2014), which includes three flood seasons (May to October).

[1]https://apps.ecmwf.int/datasets/data/tigge/

[2]http://data.cma.cn/en/?r=data/detail&dataCode=A.0012.0001/



(a) Experimental Area (b) Gridded Precipitation

Figure 5: Experimental area ($[21.0°\text{N} \sim 29.0°\text{N}, 109.5°\text{E} \sim 117.5°\text{E}]$) (left). Formatted precipitation observation (right).

We use mean interpolation to make the observation data the same format as the ensemble forecasts (as shown in Figure 5(b)). Any dates with missing observations or missing forecasts are removed, and the remaining forecast–observation pairs are quality controlled to remove unphysical values. For this study, if any one or more ensemble members are missing on a date, we remove that date from the dataset. Also, we remove data which has zero precipitation at all grids in both ensemble members and observations.

The remaining single-model dataset contained 4160 samples, of which $(16.2\%, 5.4\%)$ had greater than (10, 25) $mm \cdot day^{-1}$ precipitation, while multi-model dataset has 3785 samples.

Experimental Setup

We choose 90% samples as training data, and the other 10% as validation data. The training set and the validation set are separated in time, which is in line with the operational forecasting, and their precipitation value distribution should be consistent with the whole dataset (2 winters and 3 flood seasons). For this reason, we do not use random 10-fold crossover experiments.

We finally take the validation data from 20160416 to 20160613, the precipitation intensity distribution of which is closest to the overall data. We perform 5 runs of each setting to obtain the optimal detial structures of IC-MLNet: the network contains r residual dense structures, d dense blocks per residual dense structure, and b non-local blocks per dense block. The average of 5 outputs from the experiment with the same setting is regarded as the final performance. We enumerate $r, d, b = 1, 2, 3, 4$, and find that as long as the total number of non-local block is greater than 2, the experimental results only differ by 0.1%(RMSE)$\sim 0.7\%$(HSS) in each metric, which is approximately equal to the spread of results among repeated runs, and there is no one experiment that outperform others in all metrics. Therefore, we choose the setting with the shortest running time, i.e. $r = d = 1$, $b = 2$.

The networks are trained using Adam optimizer on a 4-GPU machine and each GPU has 4/6(Smod/Mmod) clips in a mini-batch (so in total with a mini-batch size of 16/24 clips). We train our models for 80 epochs in total, starting with a learning rate of 0.001 and reducing it by a factor of 2 at every 20 epochs. In our models, the best results are obtained when we set feature number c, c' in non-local block as 64, 16 on single-model dataset, 32, 16 on multi-model dataset. Table 1 shows more running details for Smod.

We use bias, mean absolute error (MAE), root mean squared error (RMSE) and Nash-Sutcliffe model efficiency coefficient (NSE) to evaluate the performance of various al-

GPU	Training time	Inference time
4 Tesla V100	2.19 h/run, 24 mSec/sample	3.0 mSec/sample

Table 1: Running details.

i \ j	1	2	$\cdots$	L	Total
1	$n(1,1)$	$n(1,2)$	$\cdots$	$n(1,L)$	$N_F(1)$
2	$n(2,1)$	$n(2,2)$	$\cdots$	$n(2,L)$	$N_F(2)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
L	$n(L,1)$	$n(L,2)$	$\cdots$	$n(L,L)$	$N_F(L)$
Total	$N_O(1)$	$N_O(2)$	$\cdots$	$N_O(L)$	N

Table 2: Multi-category contingency table.

gorithms:

$$Bias = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} \hat{y}_i$$

$$NSE = (1 - \sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2) \Big/ \sum_{i=1}^{n} \|\hat{y}_i - \bar{y}\|_2^2$$

where y is the prediction values (outputs, $\hat{y}$ is observed values (labels), and $\bar{y}$ denotes the average of observed values. NSE is used to quantify how well a model simulation can predict the outcome variable, ranges from negative infinity to 1, and the perfect score is 1.

We also introduce some unique evaluation indicators for precipitation categories prediction. First, we maintain a multi-category contingency table (as shown in Table 2) for each forecast map. In this table, L is the number of precipitation intensity categories (in this work, we have *None* $[0.0, 0.1)mm \cdot day^{-1}$, *Light* $[0.1, 10.1)mm \cdot day^{-1}$, *Moderate* $[10.1, 25.1)mm \cdot day^{-1}$, *Heavy* $[25.1, 50.1)mm \cdot day^{-1}$ and *Violent* $[50.1, \infty)mm \cdot day^{-1}$ five categories, i.e. $L = 5$), $n(i, j)$ denotes the number of forecasts in category i that had observations in category j ,$N_F(i)$denotes the total number of forecasts in category i, $N_O(j)$denotes the total number of observations in category j, and N is the total number of forecasts. Then we compute the evaluation indicators:

$$Acc = \frac{1}{N} \sum_{i=1}^{L} n(i,i)$$

$$HSS = \frac{\frac{1}{N} \sum_{i=1}^{L} n(i,i) - \frac{1}{N^2} \sum_{i=1}^{L} N_F(i) N_O(i)}{1 - \frac{1}{N^2} \sum_{i=1}^{L} N_F(i) N_O(i)}$$

Accuracy indicates the fraction of the forecasts in the correct category, but it is heavily influenced by the most common category. *Heidke skill score* measures the fraction of correct forecasts after eliminating those forecasts which would be correct purely due to random chance, and ranges from minus infinity to 1. 0 indicates no skill and 1 is the perfect score.

Baseline Algorithm

We compare the performance of IC-MLNet against the following four algorithms.

EM(Ensemble Mean) takes the average of all ensemble members as prediction. However, previous studies (Pan et al. 2015)(Kong, Droegemeier, and Hickmon 2006) have shown that after simple averaging, the forecast values at heavy rainfall thresholds are smoothed out, and the precipitation areas are falsely confluent.

PM(Probability Matching)(Ebert 2001) first sorts the ensemble mean and the original ensemble forecasts separately to obtain two sequences, and then segments the ensemble sequence evenly (each contains n values, n is the number of ensemble members), and finally replaces the mean sequences of corresponding segments' median to obtain a deterministic forecast. PM synthesizes the results of the ensemble forecast and usually get more accurate prediction than EM.

BP(Best Percentile)(Dai et al. 2016) is an improved PM method: the mean sequences in PM are replaced with the percentiles chosen by the forecaster, rather than the median. The BP method is more flexible, but forecasters usually can't identify the best percentile, so they tend to false alarm heavy precipitation (For details, please refer to the precipitation distribution experiment).

WEM(Weighted Bias-removed Ensemble Mean) (Zhi, Zhou, and Xu 2011) uses the deviation of forecasts and observation of training set to correct the forecasts of validation set:

$$F_{WEM} = \bar{O} + \sum_{i=1}^{n} (E_i \Big/ \sum_{i=1}^{n} E_i)(F_i - \bar{F}_i)$$

where $\bar{O}$, E_i, $\bar{F}_i$ is the mean of observations, the reciprocal of the mean error, and the mean of forecasts on the training set, and F_i is the predictions on the validation set. Furthermore, to evaluate the effectiveness of individual components of our model, we also include its several variants for the comparison: **IC-MLNet-noMS**: Multi-Supervision is not included in the basic network, **IC-MLNet-conv**: Non-local block is completely replaced by a 3×3 convolution layers, **IC-MLNet-noS**: Ensemble forecasts are not sorted before feeding into the network, **IC-MLNet-single**: There is only one layer in bias correction module, and input forecasts are not classified to different categories.

We use the same training set and validation set for all baseline algorithms as IC-MLNet. Source code is available at https://github.com/kia-kia/IC-MLNet

Results

Main Performance Analysis

Precipitation Amounts Prediction Table 3 shows that IC-MLNet achieves the best performance in terms of all precipitation amounts evaluation metrics on both datasets. Our model shows 17.8%, 35.0%, 9.7% and 75.2% improvement in terms of MAE, Bias, RMSE and NSE, respectively, on Smod compared to the best results of baseline approaches. Similarly, IC-MLNet enhances the performance by 17.9%, 37.9%, 9.2% and 71.7% on Mmod in terms of MAE, Bias, RMSE and NSE, respectively. Furthermore, we observe that the simplest EM has an excellent performance among baseline methods in the prediction of precipitation amounts, but our model surpasses it.

We evaluate the effectiveness of each individual component of IC-MLNet with an ablation study. As described in Section 4.3, each variant is different from the proposed IC-MLNet by removing one tested component. Table 3 shows that replacing the non-local block with a 3×3 convolution

	Smod						Mmod					
	Precipitation amounts				Categories		Precipitation amounts				Categories	
	MAE	Bias	RMSE	NSE	Acc	HSS	MAE	Bias	RMSE	NSE	Acc	HSS
EM	*6.1073*	1.2733	*11.5020*	*0.1971*	0.5330	0.2653	*7.1607*	1.2991	*12.5314*	*0.1891*	0.5041	0.2528
PM	6.4901	1.2794	13.2375	-0.0635	0.5616	0.3179	7.5511	1.3006	14.2221	-0.0445	0.5247	0.2865
BP	10.1254	1.5740	24.2992	-2.5834	*0.5852*	*0.3525*	8.8641	*1.2714*	18.6596	-0.7980	*0.5593*	*0.3238*
WEM	6.8066	*1.2497*	44.9405	-11.2570	0.5559	0.3407	8.9343	1.3767	39.4014	-7.0168	0.5027	0.2817
IC-MLNet-noMS	5.0587	0.8279	10.4532	0.3368	0.6007	0.3897	5.9652	0.8871	**11.3747**	0.3243	0.5623	0.3384
IC-MLNet-noS	5.0759	**0.8117**	10.4325	0.3395	0.6015	0.3862	6.0985	0.7802	11.8890	0.2631	0.5743	0.3496
IC-MLNet-conv	5.3474	0.9007	10.7197	0.3181	0.5819	0.3469	5.9330	0.8734	11.4105	0.3208	0.5756	0.3642
IC-MLNet-single	5.0870	0.8054	10.5575	0.3228	**0.6129**	0.3912	5.9914	0.8535	11.5338	0.3070	0.5451	0.2942
IC-MLNet-Mtest	-	-	-	-	-	-	6.1480	1.0032	11.8875	0.2631	0.5381	0.2920
IC-MLNet-Stest	5.3457	0.9054	11.1979	0.2541	0.5943	0.3427	-	-	-	-	-	-
IC-MLNet	**5.0185**	0.8176	**10.3863**	**0.3453**	0.6030	**0.3914**	**5.8773**	**0.7899**	11.3828	**0.3246**	**0.5784**	**0.3681**

Table 3: Performance results of precipitation amounts and intensity categories on Smod and Mmod.

will result in the biggest performance degradation both on Smod and Mmod, which shows that learning ensemble percentile contextual features is crucial to distinguish precipitation of different intensities. Moreover, in our relatively shallow network, 3×3 conv is limited in the capturing of long-range spatial features due to its small receptive field. Comparing IC-MLNet with IC-MLNet-noS shows the impact of sorting operation on the performance. We find that it brings more improvements in Mmod than Smod, but neither is prominent, indicating that the randomness of the membership order will affect the fitting ability of the network, which is consistent with previous studies. However, unexpectedly, compared with the Smod generated by random initial conditions, the member context features of the Mmod with the members ordered according to the NWP models are more difficult to capture. This may be because: first, the number of Mmod members is smaller and it is easier to be disturbed by irrelevant factors; second, the difference in forecasts generated by different NWP models is more complex to capture than that in forecasts generated by the same NWP model under random initial conditions. We find that multi-supervision and multi-layer structure provides least performance gain for both cases, yet, they also contribute to the prediction of precipitation amounts, hinting at the benefit of precipitation intensity stratified learning. Additionally, we observe that our models (variants of IC-MLNet) mostly outperform baselines in terms of precipitation value prediction, which indicates they better learn the complex spatial relationships between grids.

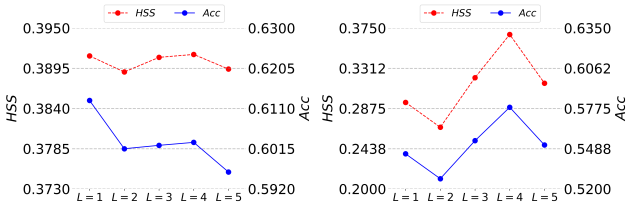
Precipitation Intensities Prediction Obviously, all algorithms (our networks and baselines) show the dilemma of accuracy in precipitation amounts prediction and intensity categories prediction: EM no longer has the advantage in the prediction of precipitation categories, the best performing method in the baselines is BP, which is unsatisfactory in amounts prediction. Nevertheless, IC-MLNet also achieves more excellent performance in predicting categories. It shows 3.0% and 11.0% improvement in terms of Accuracy and HSS, respectively, on Smod, while enhances the performance 3.4% and 13.7% on Mmod.

In addition, comparing IC-MLNet with IC-MLNet-single and IC-MLNet-noMS shows the introduction of multi-supervision and multi-layer structure makes network pays more attention to the learning of different intensity features and obtains more accurate results while correcting the central tendency. And we find that they provide less improvement on Smod compared to Mmod. That suggests inputs with fewer members benefits more from this increased dependence, because it can extract further information. Sort operation and long-range contextual dependency is still informative in both cases since their removal (IC-MLNet-noS, IC-MLNet-conv) results in obvious reduction of performance.

Statistical Test We also perform statistical test on the results with marginal improvement ($< 10\%$). Results of 5 runs of EM and BP that performed the best in baselines are collected. As shown in Table 4, we use Bayesian correlated t-test ($rope = 0.01$) to assess the significance of the achieved improvements in RMSE and Acc. Hypothesis testing shows that compared with BP and EM, IC-MLNet's improvement in RMSE is credible because $P_{(IC-MLNet\ll)} > 0.99$. Similarly, it is also obvious that IC-MLNet outperforms EM in ACC. However, the improvement to BP is not clear. Nonetheless, the probability that IC-MLNet has good performance is close to 0.99 on Smod, and the probability that it has equal performance is much smaller: $P_{(IC-MLNet=BP)} = 0.0124$ on Smod and $P_{(IC-MLNet=BP)} = 0.254$ on Mmod. This indicates

		Smod	Mmod
RMSE	$P_{(IC-MLNet\ll EM)}$	1.000	1.000
	$P_{(IC-MLNet\ll BP)}$	1.000	1.000
Acc	$P_{(IC-MLNet\gg EM)}$	1.000	1.000
	$P_{(IC-MLNet\gg BP)}$	0.987	0.746

Table 4: Results of Bayesian correlated t-test.



(a) Smod (b) Mmod

Figure 6: HSS and Acc results w.r.t change in the number of network layers L.

that we preliminarily found a postprocessing method which can maintain the accuracy of intensity categories prediction when correcting ensemble forecasts errors.

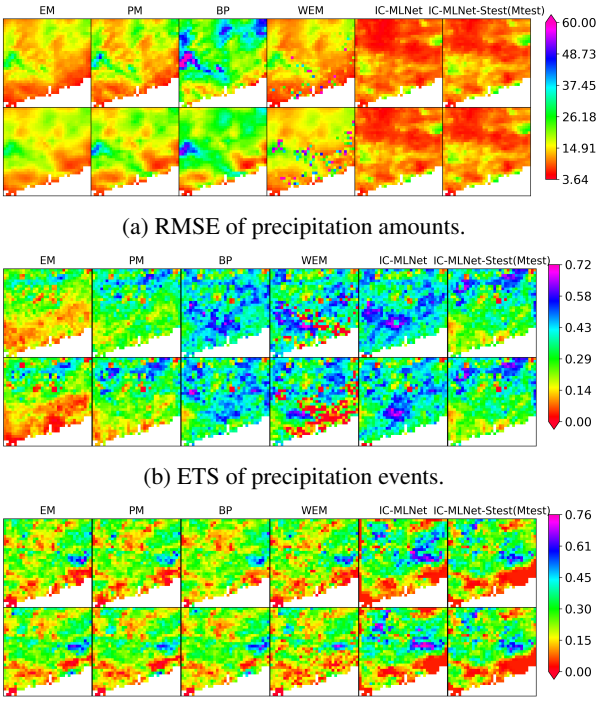
Multi-layer Analysis

We further evaluate the performance of IC-MLNet with respect to the change in number of layers L. We report results for $L \in \{1, 2, 3, 4, 5\}$ in Figure 6. When $L = 5$ add the threshold $c_5 = 100.1$. The best performances are obtained when $L = 4$ for both Smod and Mmod in terms of HSS. As L increases until the optimum value, the performance generally increases, and adding layers after reaching the best value will significantly degrade performance. These results suggest working with different intensities separately contribute to the category prediction, but if we increase the number of layers too much, it will interfere with the learning of precipitation events occurrence.

However, the best performances are obtained on Smod in terms of Acc when $L = 1$. This is because Acc is not sensitive to the false alarm rate, and often affected by the more frequent categories, and the single-layer network tends to forecast no precipitation as well as light precipitation events. Moreover, the impact of L on forecast accuracy is not as notable as those in Mmod. This situation may occur because a larger number of members is beneficial to the reduction of precipitation forecast errors, so the increased precipitation intensity feature information cannot bring obvious accuracy gain.

Precipitation Distribution Analysis

We compare the performance of IC-MLNet and baselines in the prediction of the spatial distribution of precipitation amounts. Figure 7(a) shows the RMSE of daily precipitation predicted by different methods on two data sets. The prediction results of IC-MLNet are closest to the observations. Furthermore, we note that WEM's results are visually the best of the baselines. Its notably poor overall performance in RMSE (Table 3) is attributed to some outliers, but these points can be easily excluded by the forecasters, hence it remains widely used. Our approach, which is also based on historical information (training data) but with no obvious abnormal value, goes beyond it. In addition, the mean RMSE of BP in some areas is particularly large, which is caused by its tendency to forecast heavy precipitation: the *False Alarm Ratio* (light/ moderate/ heavy/ violent) of each method on Smod is EM (0.2135/ 0.4243/ 0.6139/ 0.7930); BP (0.1129/ 0.3300/ 0.6573/ 0.8952); ICMLNet (0.0691/ 0.3127/ 0.4683/ 0.7357)



(a) RMSE of precipitation amounts.

(b) ETS of precipitation events.

(c) ETS of heavy precipitation events.

Figure 7: The precipitation spatial distribution. The first row shows results of single-model dataset and the second row shows that of multi-model dataset.

Figure 7(b) and 7(c) shows the spatial distribution of ETS of precipitation occurrence and heavy precipitation events. ETS measures the fraction of a precipitation event of specified intensity that are correctly predicted, which range from $-1/3$ to 1(larger is better). We can find that compared to baselines, the improvement of IC-MLNet is more pronounced in the prediction of yes/no precipitation than in extreme events (heavy rain). In addition, its prediction accuracy is regionally different, i.e., its performance in coastal areas is worse than that in interior areas.

Scalability of IC-MLNet

The ensemble precipitation postprocessing methods need to be extended to other datasets from different NWP models. We test scalability by computing prediction accuracy of IC-MLNet on Mmod, and it is trained on cut-Smod (IC-MLNet-Mtest). We select 4 members of Smod randomly to form cut-Smod, which is performed 5 times to reduce the deviation of the results. Also, IC-MLNet-Stest is trained on Mmod, and tested on cut-Smod. As shown in Table 3, the performance of the model is still superior to most baselines in both precipitation amounts and precipitation intensity prediction. Figure 7 shows that its spatial distribution features are also preserved. However, due to the dependence of the neural network on the training set, its accuracy has decreased compared to the

method tested on native data set (IC-MLNet).

Conclusion

Post-processing of weather forecast is a daily step that every operational center will carry out. The post-processing calibration provides the Weather Center with a lighter and more concise method that can improve the accuracy of the forecast and meet the requirements of users than updating current complicated numerical models. We proposed an ensemble precipitation forecasts postprocessing model focusing on intensity category features to generate a most likely deterministic forecast from the corresponding ensemble forecast. The multi-layer architecture was developed to correct the errors of ensemble forecasts by capturing the spatial features of different intensity categories and learning the precipitation with complex formation. Extensive experimental results on two different datasets have shown its superior performance over previous algorithms on prediction of precipitation's distribution, amounts and intensity categories and it is of good scalability.

Acknowledgments

This research was supported by the National Key R&D Program of China 2018YFB0204303, and was also supported in part by the Natural Science Foundation of China under Grant No. U1811464, No.U1811463 and No.61976250, in part by the Projects for Guangdong Introducing Innovative and Enterpreneurial Teams under Grant NO. 2016ZT06D211, in part by the Major Program of Guangdong Basic and Applied Research under Grant 2019B030302002, in part by the Guangdong Natural Science Foundation under Grant 2018B030312002. It was also supported by the CCF-Tencent Open Research Fund. We also appreciate the insightful comments and feedbacks from anonymous reviewers.

Ethical Statement

Our work helps the weather centers to achieve a higher quality forecast with low operating cost. Hereby, I consciously assure that for this paper the following is fulfilled:

1) This paper is our original work.
2) That no portion of this paper (including, but not limited to, graphics and figures) has been previously published.
3) This paper is not currently under consideration for publication elsewhere.
4) We have identified and acknowledged all sources used in the creation of our paper.
5) We have notified AAAI of any conflicts of interest we might have with regard to the work.

References

Dai, K.; Cao, Y.; Qian, Q.; Gao, S.; Zhao, S.; Chen, Y.; and Qian, C. 2016. Situation and Tendency of Operational Technologies in Short- and Medium-Range Weather Forecast. *Meteor Mon* 042(12): 1445–1455.

Ebert, E. E. 2001. Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review* 129(10): 2461–2480.

Flowerdew, J. 2012. Calibration and combi-nation of medium-range ensemble precipitation forecasts. http://www.metoffice.gov.uk/media/pdf/h/6/FRTR567.pdf. UK Met Office, Forecasting Research Technical Report 567, 2012.

Gneiting, T. 2014. *Calibration of medium-range weather forecasts*. European Centre for Medium-Range Weather Forecasts.

Hamill, T. M.; Mullen, S. L.; Snyder, C.; Toth, Z.; and Baumhefner, D. P. 2000. Ensemble forecasting in the short to medium range: Report from a workshop. *Bulletin of the American Meteorological Society* 81(11): 2653–2664.

Kalnay, E. 2003. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

Kong, F.; Droegemeier, K. K.; and Hickmon, N. L. 2006. Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part I: Comparsion of coarse-and fine-grid experiments. *Monthly weather review* 134(3): 807–833.

Krishnamurti, T.; Gnanaseelan, C.; and Chakraborty, A. 2007. Prediction of the diurnal change using a multimodel superensemble. Part I: Precipitation. *Monthly weather review* 135(10): 3613–3632.

Krishnamurti, T.; Kishtawal, C.; Shin, D.; and Williford, C. E. 2000. Improving tropical precipitation forecasts from a multianalysis superensemble. *Journal of climate* 13(23): 4217–4227.

Novak, D. R.; Bailey, C.; Brill, K. F.; Burke, P.; Hogsett, W. A.; Rausch, R.; and Schichtel, M. 2014. Precipitation and temperature forecast performance at the Weather Prediction Center. *Weather and Forecasting* 29(3): 489–504.

Pan, L.; Zhang, H.; Chen, X.; Wang, J.; and Chen, F. 2015. Neighborhood-based precipitation forecasting capability analysis of high-resolution models. *J. Trop. Meteorol* 31: 632–642.

Richardson, D. S. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* 126(563): 649–667.

Tai, Y.; Yang, J.; Liu, X.; and Xu, C. 2017. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, 4539–4547.

Wandishin, M. S.; Mullen, S. L.; Stensrud, D. J.; and Brooks, H. E. 2001. Evaluation of a short-range multimodel ensemble system. *Monthly Weather Review* 129(4): 729–747.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2472–2481.

Zhi, X.; Zhou, W.; and Xu, Z. 2011. Multi-model ensemble forecasts of the TC tracks over the Western Pacific using the TIGGE dataset. In *The 3rd international conference on information science and engi-neering. Yangzhou*, 173–4176.

Zhu, Y.; Toth, Z.; Wobus, R.; Richardson, D.; and Mylne, K. 2002. The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society* 83(1): 73–84.