Verification and Calibration of the Short-Range (0–6 h) PQPFs from Time-Lagged Multimodel Ensembles Using LAPS

Hui-Ling Chang[1,2], Huiling Yuan[3,4], Pay-Liam Lin[2] , Shu-Chih Yang[2], Yu-Chieng Liou[2], Chia-Rong Chen[1], and Wen-Ho Wang[1]

1. Meteorological Satellite Center, Central Weather Bureau, Taipei, Taiwan;
2. Department of Atmospheric Sciences, National Central University, Jhong-Li, Taiwan
3. School of Atmospheric Sciences, and Key Laboratory of Mesoscale Severe Weather / Ministry of Education, Nanjing University, Nanjing, Jiangsu, China;
4. Jiangsu Collaborative Innovation Center for Climate Change, China

Abstract

Due to the low predictability in severe weather prediction such as typhoon forecasting, it is important to develop a reliable short-range ensemble prediction system (EPS).This study aims to develop the short-range (0-6 h) probabilistic quantitative precipitation forecasts (PQPFs) of typhoons from time-lagged multimodel ensembles using the Local Analysis and Prediction System (LAPS). The ultimate goal is to provide valuable precipitation forecasts for typhoons based on the EPS.

The LAPS EPS has a good spread-skill relationship and good discriminating ability. Therefore, though it is obviously wet-biased, the forecast biases can be corrected to improve the skill of PQPFs through a linear regression (LR) calibration procedure. Sensitivity experiments for two important factors affecting calibration results are conducted, including: (1) the experiments on different training samples, and (2) the experiments on the inconsistency of observation accuracy. The first point reveals that the calibration results are sensitive to the training samples. Calibration should be performed based on consistent forecast biases between training and validation samples. The second factor indicates that the accuracy of observation is inconsistent in the sea and land areas, and samples are dominated by the ocean ones. Therefore, individual calibration for these two areas is needed to ensure better calibration results.

Keywords: probabilistic quantitative precipitation forecasts (PQPFs), verification, calibration, discriminating ability, spread-skill relationship

1. Introduction

Due to the low predictability in severe weather prediction such as typhoon forecasting, it is important to develop a reliable short-range ensemble prediction system (EPS). The EPS uses perturbed initial states or considers the physics as stochastic processes, which reflects the chaotic nature in the atmosphere. Averaging the ensemble forecasts from slightly perturbed initial conditions can filter out some unpredictable components of the forecast, and the spread among the forecasts can provide some guidance on the reliability of the forecasts. This is a fundamental transition and revolutionary change in the NWP development.

Early research indicates that calibration is a critical procedure to correct forecast biases and enhance forecast skill in a biased forecast system. In this study, the verification results showed that the LAPS EPS was apparently wet-biased. Therefore, the LR method (Yuan et al., 2008) was used to correct forecast biases.

This study (Chang et al., 2012) aims to develop the short-range probabilistic quantitative precipitation forecasts (PQPFs) of typhoons from time-lagged multimodel ensembles using the Local Analysis and Prediction System (LAPS). The ultimate goal is to provide valuable precipitation forecasts for typhoons based on the EPS. This report is organized as follows: LAPS and verified observation data are introduced in section 2. The LAPS ensemble configuration and PQPF products are presented in section 3. Section 4 and section 5 describe the verification results and sensitivity experiments on calibration. A summary is given in the last section.

2. Model and data

The short-range forecast System LAPS has three main components, including the data ingestion, diabatic data assimilation, and mesoscale model forecast (Fig. 1). In the diabatic data assimilation, the products of cloud analysis could provide the initial fields with diabatic information, such as cloud liquid water, cloud ice, and vertical motions in the cloud-covered area. Therefore, the spin-up problem could be mitigated, and the ability of short-range precipitation forecasts could be largely improved.

Regarding the observation data for precipitation verification, the radar-estimated rainfall data from the Quantitative Precipitation Estimation (QPE) and Segregation Using Multiple Sensors (QPESUMS) were used as observation data with 1.25 km horizontal resolution. Note that the precipitation estimation was calibrated with gauges in the land areas, but was not calibrated over the sea areas.

3. LAPS ensemble configuration and PQPF products

a. Time-lagged multimodel ensemble configuration

The background fields in the LAPS analysis and lateral boundary conditions are from the same sources, including the model forecasts of 1) the non-hydrostatic forecast system (NFS) at the Central Weather Bureau (CWB) and 2) the Global Forecast System (GFS) at the National Centers for Environmental Prediction (NCEP). There are two mesoscale models associated with LAPS, including the MM5 model and the WRF/ARW model. Therefore, totally four different forecast models can be used to generate multimodel ensemble forecasts with four members, including 1) LAPS-MM5:NFS (refers to LAPS-MM5 model with the background field from CWB NFS), 2) LAPS-MM5:GFS, 3) LAPS-WRF: NFS, and 4) LAPS-WRF: GFS.

In addition to the multimodel configuration, the time-lagged configuration was adopted to increase the ensemble members. The LAPS EPS (Fig. 2) has four multimodel members and each member is initialized every three hours with the forecast length of 12 hours. Thus, for the 0-6 hr ensemble precipitation forecasts, there are three time-lagged members (the 0-6 hr, 3-9 hr, and 6-12 hr QPFs) available for each multimodel member, and four multimodel members totally build up the LAPS EPS of 12 members.

b. PQPF products

The PQPFs were created based upon the precipitation forecasts from 12 members of the LAPS EPS at different thresholds. Fig. 3 shows the 0-6h PQPFs and corresponding observed probabilities at different thresholds from typhoon Fanapi, which was the most powerful typhoon to hit Taiwan in 2010 and caused a flash flood over areas of southern Taiwan. If the radar QPE is less than the selected threshold, the observed precipitation probability is zero; otherwise, it is one. In the case of typhoon Fanapi, the rainfall regions of the forecasts and observations have good correspondence.

4. Verification results

Eight typhoon cases in 2008 and 2009 (Table 1) were used to evaluate the performance of 0-6h PQPFs. Since the LAPS EPS has severe wet bias, we adopted the linear regression (LR) method to calibrate the PQPFs. In addition, the cross validation was carried out by using the cases in 2008 as the training samples to calibrate the cases in 2009, and in turn the 2009 cases as the training samples.

Some verification methods were used to evaluate the spread-skill relationship, forecast bias and discriminating ability. Regarding the experiments in this study, please refer to Table 2 for a detailed description.

a. Spread-skill relationship

A critical measure of the quality of ensemble forecasts is the spread-skill relationship. That is, whether the small-spread ensemble forecasts have smaller forecast errors (i.e., higher skill) than the large-spread ones. In this study, the ensemble spread (SPRD) was used as the spread measure, and the root mean square error (RMSE) of the ensemble mean was used as the skill measure. The scatter plot from eight typhoon cases (Fig. 4) shows that the SPRD and RMSE are highly correlated with a correlation of 0.96, which indicates a good spread-skill relationship. That is, the ensemble spread can well represent the forecast uncertainties.

b. Forecast performance

Usually, relative operating characteristic (ROC) is used together with reliability diagram to evaluate the forecast performance since ROC is conditioned on the observations and reliability diagram is conditioned on the forecasts. ROC answers the question: given that an event occurs, what is the corresponding forecast? While reliability diagram answers the question: given that an event is predicted, what is the outcome? Reliability diagram is related to conditional bias, but ROC is not sensitive to bias. A biased forecast may still have a good ROC, which means that it can be improved through a calibration procedure. Therefore, ROC can be taken as a measure of potential usefulness.

1) Discriminating ability

The relative operating characteristic (ROC；Wilks 2006) curve plots the hit rates vs. the false alarm rates using a set of probabilities thresholds (i.e., rainfall event is regarded as occurring when the forecast probability exceed this threshold). The area under the ROC curve (i.e. ROC area) measures the ability of the forecast to discriminate between events and non-events. If the ROC curve lies above the diagonal, the ROC area is greater than 0.5, which indicates skillful discriminating ability. The ROC curves (Fig. 5a) and the ROC areas (> 0.825, Fig. 5b) from the experiment SMP-T indicate good discriminating ability. It also implies that the biased PQPFs from LAPS EPS can be greatly improved through a calibration procedure.

2) Forecast bias

The reliability diagram can be used to determine how well the forecast probabilities correspond to their observed frequencies. If the reliability curve is closer to the diagonal, the forecast bias is smaller and the reliability is higher. If the curve lies below the diagonal, it indicates over-forecasting. Points above the diagonal indicate under-forecasting. Except for the slightly dry bias in lower forecast probabilities at smaller thresholds (below 5mm/6h) from the experiment SMP-T (Fig. 6a), all reliability curves indicate wet biases (i.e., the reliability curve lies below the diagonal), and the bias grows with increasing threshold. The wet biases were corrected after calibration (left column in Fig. 6).

The corresponding histogram (right column in Fig. 6) shows the sample ratio in each forecast probability bin. In general, the dry and wet biases were corrected by adjusting the lowest and highest probabilities to the mid-range ones via the calibration procedure.

5. Sensitivity experiments on calibration

In this section, two sets of sensitivity experiments (Table 2) were carried out, including (1) the experiments on different training samples, and (2) the experiments on the inconsistency of observation accuracy. The inconsistency is from the fact that the radar QPEs used as the observations were calibrated with gauges in the land area, but were not calibrated in the sea areas.

a. Experiments on different training samples

This sensitivity experiment consists of the reference (SMP-T) and its two calibration experiments, SMP-T(LR) and SMP-T8S(LR). The difference between the two calibration experiments lies in adopting different training samples during the calibration process. All samples in the experiment SMP-T(LR) were divided into two groups when carrying out the cross validation. They are typhoon cases in 2008 and 2009 respectively, where one group was used as the training samples to calibrate the other one. However, the experiment SMP-T8S(LR) puts all samples into eight groups, which are eight typhoon cases respectively. Seven groups were used as the training samples to calibrate the remaining one. In other words, each typhoon case serves as the validation samples in turn. Though the SMP-T8S(LR) adopted far more training samples, the calibration results of the SMP-T(LR) is better at the large thresholds.

Figure 7 indicates that eight typhoon cases do not show a very similar characteristic in PQPFs in terms of reliability at the threshold of 30mm/6h. Therefore, adopting more training samples [SMP-T8S(LR)] does not guarantee better calibration results. This sensitivity experiment shows that calibration results are sensitive to the training samples. Therefore, calibration should be performed based on consistent forecast biases between training and validation samples.

b. Experiments on the inconsistency of observation accuracy

This sensitivity experiment consists of two sets of experiments. The difference between the two sets lies in that their samples were adopted from different radar coverage. The samples of the experiments SMP-T and SMP-T(LR) were adopted from all radar coverage (including the sea and land areas), while those of the experiments SMP-L and SMP-L(LR) were only from the land areas in Taiwan. The land samples are only about 9.4% of all samples.

The reliability diagrams of SMP-L and SMP-L(LR) only show slightly mixed biases at each threshold. Comparing the reliability diagrams of the experiment SMP-T with SMP-L indicates that the severe wet biases in SMP-T result mainly from the ocean samples, which most likely results from the underestimation of the radar QPEs in the sea areas.

Figure 8 shows the spatial distribution of rank probability skill score (RPSS) from the experiments SMP-T and SMP-T(LR). The RPSS measures the relative improvement of the probabilistic forecast over climatology for a multi-category probabilistic forecast. The positive RPSS indicates a skillful forecast, with the perfect value of 1. After calibration, the RPSS values in most sea areas have increased, while those in a few land areas have decreased. That is, the calibration results are dominated by the ocean samples.

This is because the observation accuracy is inconsistent in the sea and land areas, and samples are dominated by the ocean ones. Therefore, individual calibration for the sea and land areas is needed to ensure better calibration results.

6. Summary

The LAPS EPS has a good spread-skill relationship and skillful discrimination ability. Therefore, the biased PQPFs can be greatly improved through a calibration procedure. The sensitivity experiments on calibration show that, first, calibration should be performed based on consistent forecast biases between training and validation samples. Second, we should consider the inconsistency of observation accuracy performing the calibration.

In the future, with more collected typhoon cases, the distributions of precipitation forecast biases can be analyzed for different typhoon paths, moving speeds or precipitation intensities. Then various LR relationships can be established and applied to different distributions of forecast biases in the typhoon cases, thus to produce better calibration results.

Reference

Chang, H. L., H. Yuan, P. L. Lin, 2012: Short-Range (0-12h) PQPFs from Time-Lagged Multimodel Ensembles Using LAPS. *Mon. Wea. Rev.*, **140**, 1496–1516.

Yuan, H., J. A. McGinley, P. J. Schultz, C. J. Anderson, and C. Lu, 2008: Short-range precipitation forecasts from time-lagged multimodel ensembles during the HMT-West-2006 campaign. *J. Hydrometeor.*, **9**, 477–491.

TABLE 1. Typhoon cases in 2008 and 2009.

	Event : number of validation time	Start — end
2008	Kalmaegi : 11 (TY1)	16 Jul, 09UTC — 18 Jul, 15UTC
	Fung-wong : 24 (TY2)	26 Jul, 03UTC — 29 Jul, 12UTC
	Sinlaku : 27 (TY3)	11 Sep, 00UTC —15 Sep, 12UTC
	Hagupit : 11 (TY4)	22 Sep, 00UTC — 23 Sep, 06UTC
	Jangmi : 17 (TY5)	27 Sep, 18UTC — 29 Sep, 18UTC
	total : 5 typhoons 90 6-h	
2009	Linfa : 12 (TY6)	20 Jun, 06UTC — 22 Jun, 03UTC
	Molave : 7 (TY7)	17 Jul, 00UTC — 17 Jul, 18UTC
	Morakot : 39 (TY8)	05 Aug, 18UTC — 10 Aug, 2UTC
	total : 3 typhoons 58 6-h	

TABLE 2. Summary of the difference of statistical samples in the sensitivity experiments.

Expt	Description
SMP-T	Statistical samples were adopted from all radar coverage area within the QPESUMS domain (including sea and land areas), before LR calibration. This experiment was used as a reference one in this study.
SMP-T(LR)	Statistical samples were the same as in the experiment SMP-T, but after LR calibration. The statistical samples were divided into two groups (cases in 2008 and 2009 respectively) when performing the cross-validation procedure.
SMP-T8S(LR)	The same as in the experiment SMP-T(LR), but the statistical samples were divided into eight groups (eight different typhoon cases in 2008 and 2009) when performing the cross-validation procedure.
SMP-L	The same as in the experiment SMP-T, but the statistical samples were only adopted from the land area within the QPESUMS domain.
SMP-L(LR)	The same as in the experiment SMP-T(LR), but the statistical samples were only adopted from the land area within the QPESUMS domain.

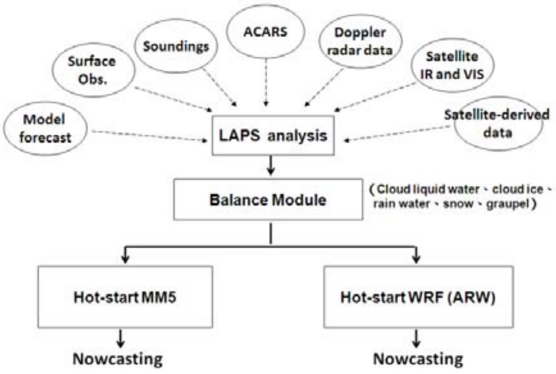

FIG. 1. Schematic diagram of short-range forecast system LAPS.

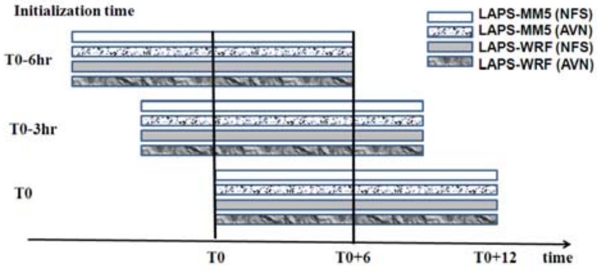

FIG. 2. Schematic diagram of time-lagged multimodel ensemble.

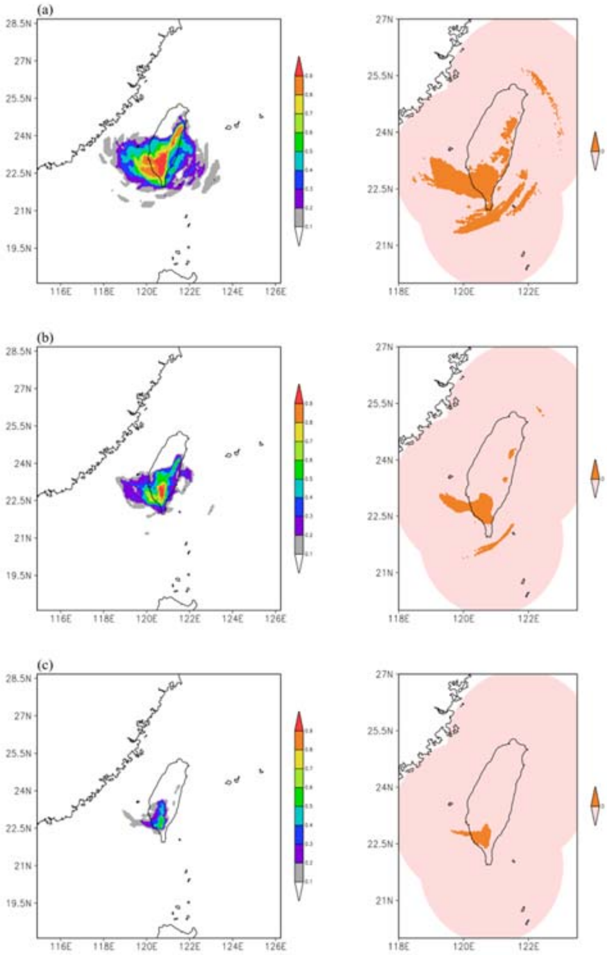

FIG. 3. Distribution of LAPS 0-6 h PQPFs (left column) and QPESUMS precipitation (used as truth) probabilities (right column) at thresholds (a) 50, (b) 100, and (c) 200 mm/6 h ending at 1200UTC 19 Sep 2010. In the right column, orange shaded area denotes pixels where QPESUMS precipitation estimations exceed the indicated threshold, and pink shaded area indicates QPESUMS radar coverage.

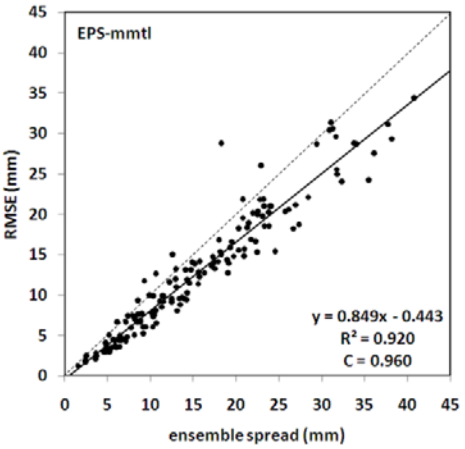

FIG. 4. Scatter plots of the RMSE against the ensemble spread (SPRD) from the experiment SMP-T. Each point in the scatter plot comes from one 0-6 h QPF (i.e. RMSE and SPRD are averaged over the QPESUMS domain). The linear regression line, correlation coefficient (C), and the coefficient of determination (R^2) are shown on the plot.

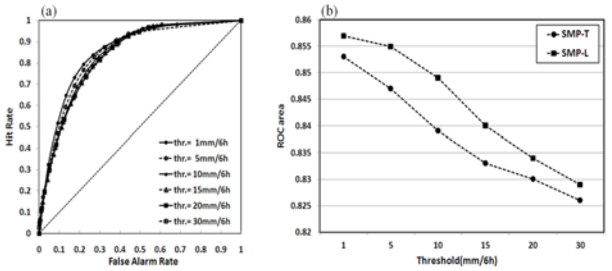

FIG. 5. (a) ROC curves from the experiment SMP-T and (b) the area under the ROC from the experiments SMP-T (line with circles) and SMP-L (line with squares) for LAPS 0-6 h PQPFs at different thresholds (1, 5, 10, 15, 20, and 30 mm/6 h).

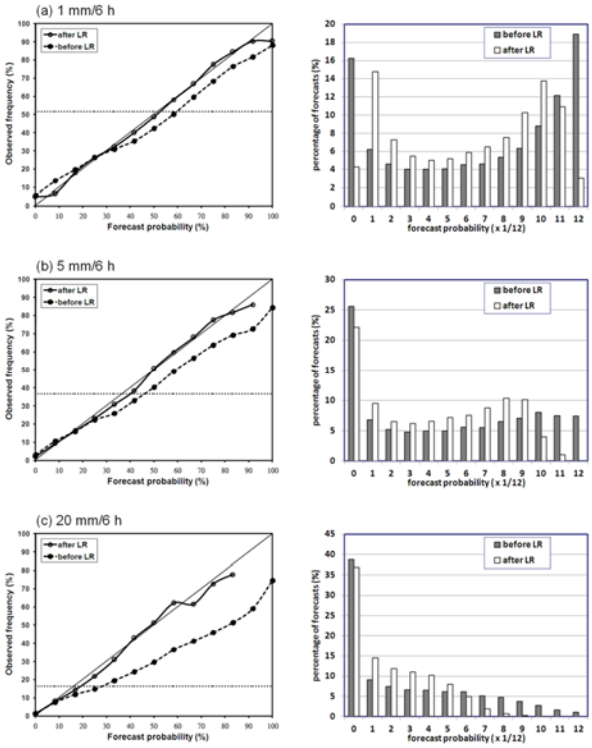

FIG. 6. Reliability diagrams (left) for LAPS 0-6 h PQPFs at thresholds (a) 1, (b) 5, and (c) 20 mm/6 h. Reliability curves from the experiments SMP-T (before LR calibration, dashed line with solid dots) and SMP-T(LR) (after LR calibration, solid line with hollow circles) are shown. The horizontal dashed line indicates the sample climatology frequency. Histograms (right) indicate the corresponding sample ratio (%) of each forecast probability subrange for the experiments SMP-T (before LR, gray) and SMP-T(LR) (after LR, blank).

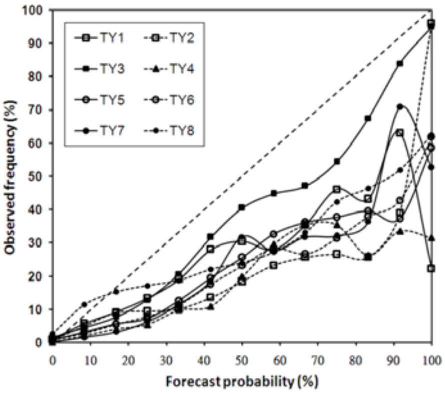

FIG. 7. Reliability diagram for LAPS 0-6 h PQPFs at the threshold of 30 mm/6h. Reliability curves from eight typhoon cases (TY 1 to TY 8) in 2008 and 2009 are shown.

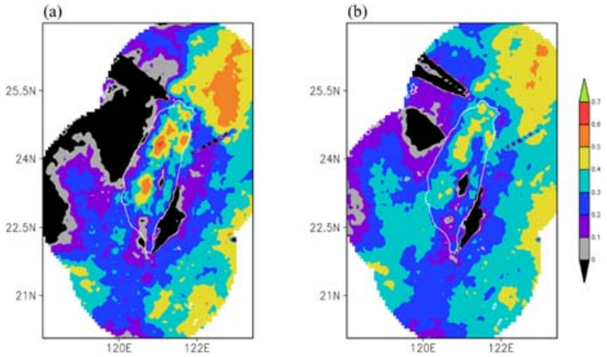

FIG. 8 The spatial distribution of the ranked probabilistic skill score for LAPS 0-6 h PQPFs from the experiments (a) SMP-T (before LR calibration) and (b) SMP-T (LR) (after LR calibration) using four thresholds (1, 5, 10 and 20 mm/6 h) to define five categories.