Are the discrimination ability and maximum economic value of a forecast system affected by forecast bias?

Hui-Ling Chang[1] and Shu-Chih Yang[2]

1. Meteorological Satellite Center, Central Weather Bureau, Taipei, Taiwan;
2. Department of Atmospheric Sciences, National Central University, Jhong-Li, Taiwan

Abstract

Most forecast systems possess systematic biases because of the crude representation of model physics and dynamics to the real atmosphere. Focusing on short-range probabilistic quantitative precipitation forecasts (PQPFs) for typhoons, this study explores the effect of calibration on the discrimination ability and maximum economic value (EV_{max}) based on the Local Analysis and Prediction System (LAPS) ensemble prediction system (EPS) operated at Central Weather Bureau in Taiwan.

Results show that the discrimination ability, which can be assessed by relative operating characteristic (ROC), and EV_{max} of an EPS are insensitive to forecast bias. This implies that improving reliability via calibration cannot increase the discrimination and EV_{max} of a forecast system. However, the optimal probability threshold (P_t) for users to take preventive action and obtain their EV_{max} is different when adopting calibrated or uncalibrated forecasts. In other words, biased forecasts will not prevent users from obtaining their EV_{max} if an appropriate Pt is adopted. When uncalibrated forecasts are adopted, the optimal P_t should be determined based on the past long-term statistics of EV.

Experiments have been conducted to verify that ROC is insensitive to the linear property of a calibration method or even the accuracy of the calibration results. That is, the discrimination ability almost remains the same after a linear or nonlinear calibration, even though forecast biases cannot be properly corrected during the calibration process. This is because calibration only corrects the precipitation amount instead of modifying the rainfall pattern, which is controlled by the model physics/dynamics processes and is associated more with the discrimination of the forecasts.

Keywords: probabilistic quantitative precipitation forecasts (PQPFs), relative operating characteristic (ROC), economic value (EV)

1. Introduction

The quality of an ensemble prediction system (EPS) can be quantitatively assessed by a host of verification metrics, such as the spread-skill relationship, the rank histogram, the reliability diagram, the relative operating characteristic (ROC), the Brier skill score (BSS), and the rank probability skill score (RPSS). Different verification methods are designed to assess different characteristics (or attributes) that contribute to the quality of a forecast, such as accuracy, reliability, discrimination, skill, and so forth. The discrimination (i.e., the ability to discriminate between events and non-events) of ensemble probabilistic forecasts (EPFs) can be measured by the ROC, which is based on the signal detection theory (SDT; Harvey et al. 1992) and is widely used in the fields of economy and social sciences, such as psychology and medicine, and was introduced to meteorology by Mason (1982). Unlike reliability diagrams conditioned on the forecasts, the ROC is conditioned on the observations. Therefore, the ROC is a good partner to the reliability diagram.

In addition to the forecast quality, one should also consider the possible economic benefit in the daily decision-making process of users when measuring the usefulness of weather forecasts. In this study, we used the relative economic value (EV; Richardson 2000) to assess the economic benefit of users, which is defined as the reduction in expected expense over the use of purely climatological information relative to the reduction that would be obtained by using perfect forecast. Murphy (1977) showed that for a perfectly reliable forecast system (i.e., without forecast bias), users can obtained the

maximum EV (EV_{max}) if they adopted the probability threshold (P_t) equal to their cost–loss ratio (r). The choice of P_t converts the EPF to a deterministic forecast, which treats the forecast with the probability greater than P_t as the occurring event, and users can take preventive actions based on the P_t, such as shutting down roads, harvesting crops in advance, and suspending work and school.

Several studies (Richardson 2001; Zhu et al. 2002) indicate that the ROC is closely related to the EV, and Chang et al. 2014 shows that the potential EV (the area under the curve of EV_{max}) provided by a forecast system is mainly determined by the discrimination ability of the same system. In addition, Chang et al. 2012 shows that a biased forecast system, such as the LAPS EPS, can still have good discrimination. Based on these interesting results, this study focuses on the effect of calibration on the ROC and EV_{max} since most forecast systems have systematic biases and need calibration before being used by users. This purpose of this study is to explore whether a calibration procedure could improve the discrimination ability of a forecast system and thus increase the EV_{max}.

This paper is organized as follows: Section 2 introduces the Local Analysis and Prediction System (LAPS) ensemble prediction system (EPS) and data. Section 3 describes the methodology for computing ROC and EV, as well as the linear regression method used to calibrate the forecast bias of LAPS EPS. Section 4 presents the effects of calibration on ROC and EV_{max}. The sensitivity of ROC curve to the form of calibration are also discussed in Section 5. Finally, a summary is provided in Section 6.

2. LAPS EPS and data

The 0-6 h probabilistic quantitative precipitation forecasts (PQPFs) used in this study were generated from ensemble forecasts based on the LAPS. By adopting diabatic data assimilation, the LAPS mitigated the spin-up problem and performed reasonable precipitation forecasts during the early stage of a forecast period. The LAPS PQPFs were operationally generated every 3 h at the Central Weather Bureau (CWB) in Taiwan. To produce four basic multi-model ensemble members, the LAPS EPS adopted two types of background states, including the forecasts of the Global Forecast System (GFS) at the National Centers for Environmental Prediction (NECP) and the non-hydrostatic forecast system (NFS) at the CWB, to construct two sets of analysis fields with a horizontal resolution of 9 km, and then initializes two mesoscale models, including the MM5 and WRF/ARW models. For each of these basic members, the EPS adopted two more members initialized from the analyses generated three and six hours earlier. Therefore, in total, 12 time-lagged multi-model members are available and the PQPFs used in this study are derived from this 12-member LAPS EPS. Chang et al. (2012) showed that the LAPS EPS has a good spread-skill relationship and skillful discrimination ability, and thus can be regarded as an EPS with good quality and predictive capability. The data used in this study (same as in Chang et al. 2012) for evaluating the ROC and EV include a total of 148 cases of 0–6 h PQPFs based on all typhoon cases in 2008 and 2009.

A calibration method based on linear regression (Yuan et al. 2008) has been used to calibrate the wet-biased PQPFs. Chang et al. (2012) show that this calibration method successfully corrects the wet bias and improves the post-processing forecast skill.

3. Methodology

a. Relative operating characteristic (ROC)

The ROC has been widely used in meteorology to study the potential usefulness of a forecast system. The ROC was derived from the SDT (Harvey et al. 1992), which asserts that the uncertainty of the occurrence and non-occurrence of an event can be described by the relative variation of two Gaussian probability distributions: one (i.e., the event or signal distribution) represents the probability distribution of evidence strength associated with the occurrence of the event; the other (i.e., the nonevent or noise distribution) represents the probability distribution of evidence strength associated with the non-occurrence of the event. The occurred and non-occurred observation and forecast events are usually summarized by a 2×2 contingency table (Table 1). The forecast performance can be categorized based on the relative frequencies of four different outcomes: the hit (h), miss (m), false alarm (f) and correct rejection (c) and $h + m + f + c = 1$. Given the criterion of a decision, two independent conditional probabilities can be generated based on Table 1: the hit rate ($HR=h/(h+m)$) is the probability of predicting an event given that the event occurs, and the false alarm rate ($FAR=f/(f+c)$) is the probability of predicting an event given that the event does not occur.

The ROC curve for the EPFs is constructed using N pairs of (HR, FAR) for N P_t values by plotting HR (in the

y-axis) against *FAR* (in the x-axis). If the (*HR*, *FAR*) points bunch in the upper-left corner of the ROC plot (i.e., a large *HR* and small *FAR*), the events and non-events can be clearly distinguished, i.e., the forecast system has good discrimination for both events and nonevents. If the ROC plot is diagonal-dominated, the values of *HR* and *FAR* are comparable, suggesting that the probability distributions of events and non-events almost overlap and cannot be discerned. Based on this SDT idea, the area under the ROC curve, called the ROC area, is used to represent the ability of the forecasts to discriminate between events and nonevents. The ROC area ranges from 0 to 1, where 1 indicates a perfect forecast. Forecasts with skillful discriminating ability have ROC areas greater than 0.7 (Buizza et al. 1999), and the forecasts adopting climatology are unskillful with the ROC area of 0.5. Further details can be found in the references (Wilks 2006).

b. Economic Value (EV)

The EV of a forecast system (Richardson 2000) is defined as:

$$EV = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}}. \quad (1)$$

where $E_{climate}$, $E_{forecast}$ and $E_{perfect}$ are the expected expenses of a user who takes preventive action based on the climatological information, a forecast system, and a perfect deterministic forecast system, respectively. According to the above definition, the EV can be interpreted as the relative performance taking the climatological information as a baseline. For example, if a perfect forecast can save the user 100 dollars, then a forecast system with economic value EV will save the user 100×EV dollars. Richardson (2000) further showed that EV can be expressed as:

$$EV = \frac{\min[\bar{o}, r] - FARr(1-\bar{o}) + HR\bar{o}(1-r) - \bar{o}}{\min[\bar{o}, r] - \bar{o}r}. \quad (2)$$

Equation (2) shows that EV is related not only to the *FAR* and *HR* of a forecast system but also to the climatological frequency ($\bar{o}$) of a weather event and the cost-loss ratio (*r*) of a user. Since ROC is defined by *FAR* and *HR*, eq. (2) also indicates that EV and ROC are associated (Zhu et al. 2002).

c. Linear regression (LR) method

Following Yuan et al. (2008), the LR method is used for calibrating LAPS PQPFs. The LR equation is expressed as:

$$P(x,t) = a + \sum_{i=1}^{M} b_i f_i(x,t). \quad (3)$$

where M=7, $f_i(x,t)$, i= 1, 2, …, 7 are the seven ordinal ensemble precipitation probabilities centered at the calibration threshold, $P(x,t)$ is the corresponding observed precipitation probability, and a is the error residual. The LR method successfully corrected the wet bias of LAPS PQPFs and improved forecast reliability and skill (Chang et al. 2012).

4. Effect of calibration on ROC and maximum EV

Most forecast systems possess systematic biases because of the incompleteness of model physics and dynamics. The LAPS PQPFs have an obvious wet bias, and the bias becomes more apparent with the increasing precipitation intensity (Chang et al. 2012). This section explores the effect of calibration on the ROC and EV of a forecast system.

Figure 1a shows that the ROC curves before and after calibration at the 20 mm $(6\ h)^{-1}$ precipitation threshold are very similar. During the calibration process, the forecast probability (P_f) was adjusted to yield superior consistency between the P_f and the observed frequency (i.e., improving the reliability). However, the discrimination of the EPS has not been improved through the calibration procedure. Therefore, the (*HR*, *FAR*) points corresponding to the calibrated P_f on the ROC curve only shift along the curve derived from uncalibrated PQPFs. In other words, with or without the calibration, the LAPS PQPFs possess the same discrimination ability.

Figure 1b compares the distribution of EV_{max} from the LAPS PQPFs with and without calibration and also the optimal P_t that users with different *r* must adopt to achieve their EV_{max}. If an appropriate P_t is adopted, the EV_{max} from the LAPS PQPFs before and after calibration almost did not change. However, if adopting un-calibrated PQPFs, the chosen optimal P_t is higher than the theoretical value due to the wet bias of LAPS PQPFs. This will prevent users from achieving their EV_{max} if they adopt the theoretical P_t as their optimal P_t. For example, the theoretical P_t for users with r = 5/12 should be 5/12;

however, when un-calibrated PQPFs are adopted, prevention is only necessary when $P_f \geq 9/12$ because of the wet bias. Therefore, taking action at $P_f = 5/12$ (over confidence in the forecast) will result in over-prevention, which wastes money and reduces EV. Therefore, when un-calibrated PQPFs are adopted, the optimal P_t must be determined based on the past long-term statistics of economic value.

The analyses of ROC and EV also confirm that reliability and discrimination are two independent statistical characteristics to describe the performance of a forecast system. Although a forecast system may achieve perfect reliability through calibration, its discrimination, which is reflected in the ROC area and EV_{max}, is not affected by calibration. This is expectable since calibration to improve the reliability of a forecast system only corrects the precipitation amount instead of modifying the rainfall pattern, which is controlled by the model physics/dynamics processes and is more associated with the discrimination of the forecasts. Although ROC and EV_{max} are both insensitive to forecast bias, a systematic bias causes the real optimal P_t to deviate from the theoretical one. Therefore, directly using a theoretical optimal P_t for the biased (i.e., not perfectly reliable) ensemble probabilistic forecasts (EPFs) will implicitly increase the $E_{forecast}$, and thus prevent the users to reach their EV_{max}.

5. Sensitivity of ROC curve to the form of calibration

In this section, we explore the sensitivity of the ROC curve to the form of the calibration. The purpose of the sensitivity test is to confirm that the discrimination ability, measured by the ROC area, is a potential characteristic of a forecast system and cannot be modified via the calibration process. Three sensitivity experiments are conducted: (1) After the eight regression coefficients are derived, the coefficient for the ensemble probability at the calibration threshold is reset to 0 (i.e., $b_4 = 0$). (2) Instead of the LR equation, a nonlinear equation is used:

$$P(x,t) = a + b_1[f_1(x,t)]^2 + \sum_{i=2}^{6} b_i f_i(x,t) + b_7[f_7(x,t)]^2 \quad (4)$$

Because the values of the coefficients b_1 and b_7 are larger than the others, we modified the LR equation to the nonlinear equation [eq. (4)] with quadratic terms at the first and seventh thresholds. (3) After the eight coefficients in eq. (3) are derived, the error residual term is reset to 0 (i.e., $a = 0$). This has a substantial effect on the correction of biased forecasts.

It is shown that the ROC curves before and after calibration overlap in all the three sensitivity experiments (Fig. 2), Therefore, the ROC areas almost remain the same after a linear or nonlinear calibration, even the forecast biases cannot be properly corrected during the calibration process (such as in sensitivity experiments (1) and (3)). The result shows that the discrimination ability of a forecast system is insensitive to the reliability of the same system.

6. Summary

Most forecast systems possess systematic biases because of the incompleteness of model physics and dynamics. Focusing on short-range PQPFs for typhoons, this study explores the effect of calibration on the discrimination ability and EV_{max} based on the LAPS EPS operated at Central Weather Bureau in Taiwan.

Results show that discrimination ability and EV_{max} of a forecast system are insensitive to forecast bias. Calibration, though improving the reliability, has no effect on increasing the discrimination and EV_{max} of a forecast system. In addition, the biased forecast with imperfect reliability will not prevent users from achieving their EV_{max} if the appropriate P_t can be adopted. Because of the significant dominant wet bias in the LAPS EPS, the real optimal P_t for users to achieve their EV_{max} deviates from the theoretical value, which is equal to his/her *r*. Such deviation increases largely as the rainfall intensity increases, because the wet bias of LAPS PQPFs becomes more obvious when the rainfall threshold is larger. Nevertheless, by adopting the real optimal P_t, the users can still obtain the same EV_{max} by referencing the LAPS PQPFs either with or without bias correction. When uncalibrated EPFs are adopted, the optimal P_t should be determined based on the past long-term statistics of EV in order to achieve EV_{max}.

Experiments have also been conducted to verify that ROC is insensitive to the linear property of a calibration method or even the accuracy of the calibration results. That is, the discrimination ability almost remains the same after a linear or nonlinear calibration, even though forecast biases cannot be properly corrected during the calibration process. This is because calibration only corrects the precipitation amount instead of modifying the rainfall pattern, which is controlled by the model

physics/dynamics processes and is associated more with the discrimination of the forecasts.

Reference

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168–189.

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.

——, S.-C. Yang, H. Yuan, P. L. Lin and Y. C. Liou, 2015: Analysis of relative operating characteristic and economic value using the LAPS ensemble prediction system in Taiwan area. *Mon. Wea. Rev.*, **143**, 1833–1848.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883.

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, 30, 291–303.

Murphy, A.H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **126**, 649–667.

——, 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

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.

Zhu, Y., Z. Toth, R. Wobus, D. S., Richardson, and K. Mylne, 2002：The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.

TABLE 1. Contingency table for forecasts and observations of a binary event.

		Forecast / action	
		Yes	No
Observation	Yes	Hit (*h*) Mitigated loss (C+L_u)	Miss (*m*) Loss (L_p+L_u)
	No	False alarm (*f*) Cost (C)	Correct rejection (*c*) No cost (N)

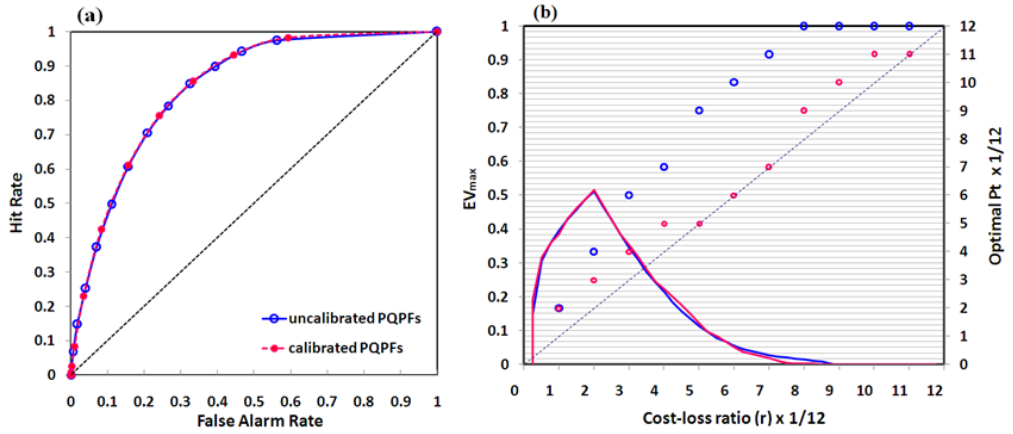

FIG. 1. (a) The ROC curves and (b) maximum economic value (EV_{max}; curves) and optimal probability threshold (P_t; circles) against the cost-loss ratio (r) before (blue) and after (pink) calibration at the 20 mm (6 h)$^{-1}$ precipitation threshold.

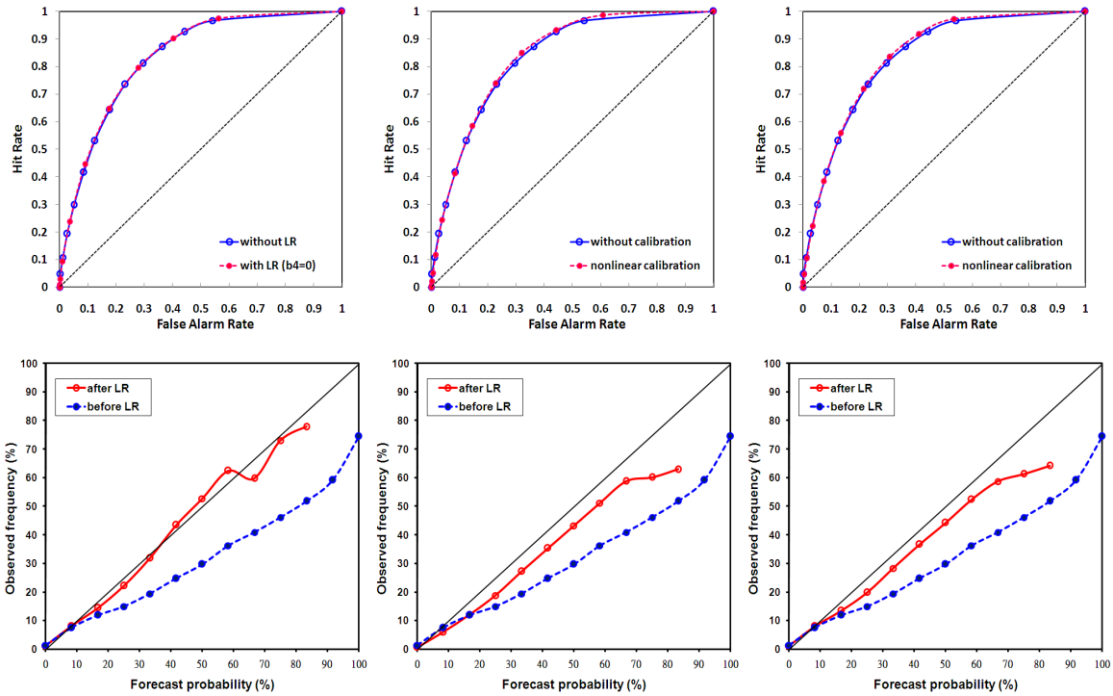

FIG. 2. The ROC and reliability curves before (blue curves) and after calibration (red curves) at the 20 mm (6 h)$^{-1}$ threshold. The b_4 is reset to zero after the coefficients are derived in the LR equation (left column), adopting the nonlinear calibration method [Eq. (4); middle column] and the error residual term is reset to 0 (i.e., a = 0) after the coefficients in Eq. (3) are derived (right column).