香港天文台
HONG KONG OBSERVATORY

Reprint 1356

Calibration of Medium-range Quantitative Precipitation Forecast (QPF) Using Spatial Statistics of High-resolution Models

Man-Lok CHONG and Wai-kin WONG

The 33rd Guangdong - Hong Kong - Macao Seminar
on Meteorological Science and Technology
and
The 24th Guangdong - Hong Kong - Macao Meeting
on Cooperation in Meteorological Operations

(Hong Kong 6-8 March 2019)

使用高分辨率数值预报模式的空间统计数据校准中期定量降水预报

庄民诺　黄伟健

香港天文台

摘要： 由于不同数值预报模式的模式气候不一，表现亦各异，定量降水预报一向是天气预报其中一项最具挑战性的工作。随着全球模式的分辨率不断提升，越来越多模式网格点覆盖到香港及邻近地区。一直以来，定点定量降水预报都是透过简单降尺度方法得出。在这个研究，我们尝试使用网格点的空间统计数据校准中期定量降水预报。参考多个全球模式在香港及附近的网格点，我们对实测小雨(0.05 毫米/日)、中雨(10 毫米/日)及大雨(25 毫米/日)为下限进行校准。校准过程使用了两种方法将训练集的临界成功指数(CSI)最大化，分别是所有网格点的平均值，以及网格点按不同阀值和不同百分比组成的二维热图。结果发现，前者比较适用于小雨，后者则较适用于中雨和大雨。验证集的 CSI 一般较同期的模式直接输出数据为高，在小雨和大雨尤其显著。

Calibration of Medium-range Quantitative Precipitation Forecast (QPF) Using Spatial Statistics of High-resolution Models

Man-Lok CHONG Wai-kin WONG

Hong Kong Observatory

***Abstract*：** Forecasting of precipitation is one of the most challenging tasks in weather forecasting, owing to different performance of Numerical Weather Prediction (NWP) models and their differences in model climate. With increasingly higher resolution grids available from global NWP models, more and more grid points cover Hong Kong and its neighbouring regions now. Traditionally, quantitative precipitation forecasts (QPF) at specific location are obtained through simple downscaling from direct model output (DMO). In this study, we attempt to calibrate medium-range QPF by making use of spatial statistics over the grids. Based on grid points from several global models in or near Hong Kong, actual rainfall with thresholds of light (0.05 mm), medium (10 mm) and heavy (25 mm) intensity has been calibrated. The calibration optimises the critical success index (CSI) of the training set by two approaches, either by the mean of all grid points, or percentage of grid points with rainfall above various thresholds, which is a two-dimensional heat map. It is found that the former method is more effective for predicting light rain while the latter more effective for forecasting of medium and heavy rain. In the verification set, the CSI for forecasts after calibration generally improved over DMO of the same period, especially for light and heavy rain.

1. Introduction

Precipitation forecast is challenging not only due to stochastic nature of underlying weather processes and highly skewed distribution (Shrestha *et al.*, 2015), but also differences in performance of numerical weather prediction (NWP) models as a result of data assimilation techniques, model resolution, formulation of model dynamics and physics. NWP models have their own characteristics in quantitative precipitation forecast (QPF), thus calibrations are often required to correct direct model output (DMO), or to derive more useful guidance based on verification of DMO against the rainfall observations (Duan *et al*., 2017).

As NWP models have different characteristics of QPF, forecast errors and uncertainties, it is more desirable to adopt a consensus or probabilistic approach to optimize the skill of rainfall forecast from available NWP models. To transform model QPF to a probabilistic forecast, Sloughter *et al.* (2007) discussed the use of logistic regression to differentiate "rain" or "non-rain" cases, followed by applying the Bayesian model averaging (BMA) method to form a probability density distribution of the daily precipitation forecast. In order to remove biases from QPF from the ensemble prediction system, Brown and Seo (2010) used the real-time ensemble forecast as a condition to generate a bias-free cumulative distribution function of precipitation, without assuming the distribution of ensemble QPF from NWP models. Another method named "Schaake shuffle" matches historical precipitation events from similar dates among a group of stations within a sufficiently large geographical domain with the ranked model ensemble members, so that the spatial correlation of precipitation in the domain can be preserved and the bias can be removed in the calibrated ensemble members (Clark *et al.*, 2004; Shrestha *et al.*, 2015).

In recent years, attempts have been made to calibrate short-range QPF from deterministic model outputs or guidance products. In the Japan Meteorological Agency (JMA), Kalman filter technique using predictors such as model prognostic wind on 850 hPa level and column moisture was developed to reduce the root-mean-square error (RMSE) of the precipitation forecast (JMA, 2013). Since model output often underestimates rainfall forecast, in particular during significant convective or extreme weather events, an adaptive frequency matching method has been used to post-process the model rainfall forecasts. A similar frequency matching method was also developed in HKO to calibrate the QPF for Hong Kong from global and regional NWP models (Li *et al.*, 2014).

As the grid resolution from global NWP models have been increasingly fine, precipitation patterns are better resolved with more available grid points covering Hong Kong and the nearby areas. This can improve the representativeness of precipitation field, especially the significant rainfall due to mesoscale weather processes, that has been smoothed or under-forecast substantially in lower resolution models in the past. However, using single grid point of high-resolution models has a risk of jumpy forecasts of precipitation due to position and intensity errors. The error of QPF cannot be easily reduced through statistical post-processing for single grid point even with the use of ensemble prediction system (EPS). Considering the spatial variability of precipitation systems from high-resolution NWP models, it would potentially improve categorical QPF guidance for Hong Kong (light, medium and heavy, etc.) if spatial characteristics of accumulated precipitation can be employed to determine suitable threshold of rainfall intensity and proportion of grid-point rainfall exceeding the threshold to optimize the skill of model QPF.

2. Data and methodology

This study aims at developing a new "calibrated" QPF guidance based on available highest-resolution NWP model data to optimize the skills of daily rainfall forecasts in Hong Kong (see the following two paragraphs). The daily rainfall forecast from the deterministic models of the European Centre for Medium-Range Weather Forecasts (ECMWF), the United States National Centers for Environmental Prediction (NCEP), the Met Office of the United Kingdom (UKMO) and the JMA. The grid points within the domain of 22.0-22.6°N and 113.7-114.5°E (Figure 1) are extracted from the models for the period 1 April 2017 to 30 September 2018, data in the first 12 months are taken as the training data set and those in the latter 6 months constitute the verification data set. In considering the available forecast time steps in DMO of the above models, the daily rainfall of a given day starts from 18Z and ends at 18Z on the following day. "Day 1" forecast from the 12Z run and 00Z run is taken to start at T+6 h and T+18 h respectively.

Mesoscale processes associated with severe weather systems and effect of orography in Hong Kong can lead to large spatial variation in accumulated precipitation over the territory (https://www.weather.gov.hk/cis/climahk_e.htm). Currently at the Hong Kong Observatory (HKO), the daily rainfall forecasts of both subjective (forecasters) and objective forecasts (e.g. model QPFs) are verified against an actual daily accumulated rainfall that is calculated as a simple average of

measurements from selected rain gauges. In this study, gridded daily QPFs over the domain (Figure 1) from the four NWP models (ECMWF, NCEP, JMA and UKMO) on each forecast day are checked against the actual daily rainfall with thresholds of light (0.1 mm or above), medium (10 mm or above) and heavy (25 mm or above) categories to determine respective QPF thresholds (light / medium / heavy) with optimized skill scores. These QPF thresholds will then be used as the "calibrated" rainfall forecasts of the NWP models for the three categories above. Table 1 shows the data resolutions, number of grid points and available forecast range.

The critical success index (CSI; see Appendix) (Donaldson *et al.*, 1975) is a common measure to show the fraction of observed and/or forecast events that are correctly predicted. The calibration optimises CSI of the forecasts in the training period using the following two approaches:

Method 1 - using mean QPF of each model over all grid points. CSI is calculated across different mean values where QPF mean with the highest CSI is picked as the optimal threshold;

Method 2 - considering performance of QPF in two perspectives – rainfall amount and percentage of grids with rainfall above the respective rainfall threshold. Various combinations of the values of the two criteria will be examined to derive an optimal combination of thresholds that maximizes CSI in the training period. Since there are two degree of freedom, results are presented using a heat map for easier understanding. The calculation of CSI is repeated for each model, each forecast run (00Z and 12Z) and each forecast day against the thresholds of the three rainfall categories.

3. Effectiveness of the two methods on different thresholds

To find out whether thresholding with a simple mean or using distribution of intensity of grid points is the more effective method for light, medium and heavy categories, both methods are applied to Day1 - Day 3 datasets of all model runs, as later days are expected to have a weaker and noisier signal.

Figures 2(a)-(c) are examples of heat maps of the four NWP models showing the CSI of different thresholds. Common data set is adopted such that the same sets of model runs are used. In the heat maps (upper panels), each grid represents a potential criterion of "calibrated QPF" in forecast of daily rainfall in HK

based on the percentage of grids (y-axis) above a certain intensity of precipitation (x-axis) during the training period. The values of CSI to the top right and bottom left are usually lower, because the criteria are either too stringent or too loose. ECMWF and UKMO have better skills in capturing precipitation of each rainfall category, as they have higher maximum CSI across different combination of forecast rainfall amount and percentage of grids. The observation is in good agreement with subjective assessment of forecasters that ECMWF and UKMO models have better performance in rainfall forecast in general.

Table 2 summarises the differences between the average maximum CSI from Method 1 and that of Method 2 to compare the effectiveness of the two approaches. All forecast days and rainfall thresholds have positive differences in CSI, meaning that Method 2 is more effective. One might intuitively pick from the heat map with the highest maximum CSI in the training process as the most optimal "calibrated" model QPF. However, it can be seen that the improvement of maximum CSI for light rain is not significant using Method 2, and simply using the mean across the grid points as threshold is sufficient to yield a good "calibrated" QPF. In forecasting medium and heavy rain, Method 2 has a more significant gain in the maximum CSI and it would be a preferable approach to be used in operation.

4. Interpretations of NWP model QPF

Based on the discussion in Section 3, Method 1 with simple mean across grid points will be applied for forecasting light rain, while Method 2 for forecasting medium and heavy rain. Figures 3(a)-(c) show the maximum CSI, which correspond to the optimal threshold, for each NWP model in the training set calibrated against thresholds of light, medium and heavy rain. The forecast skill decreases with forecasting days as the uncertainty of weather forecast increases. The higher the category of rain to be calibrated, the steeper the deteriorating trend with forecast days. Based on the CSI of different forecast days, ECMWF is the best forecast model for light rain or above while both ECMWF and UKMO have edges over the other models in medium and heavy rain forecasts.

It is observed that CSI for light rain or above remains well above 50% even for the seventh or eighth forecast day for all models. From actual records, there are 154, 54 and 26 days with precipitation above 0.1 mm, 10 mm and 25 mm respectively. The precipitation forecast (of what category??) for the first two days are not much more skilful than medium rain of the same days.

In addition to the maximum CSI determined from the heat maps, the patterns of CSI distribution across different rainfall thresholds and percentage of grids provide insights in the use and interpretation of gridded QPF. Figures 4 to 7 show the heat maps of ECMWF, UKMO, NCEP and JMA 12Z runs for their available forecast periods respectively. Differences of CSI distributions are obvious comparing the best two models with the worst two where yellowish CSI values are not present. In particular, JMA model shows rapid deterioration of forecast skills in the last few days. Grossly comparing the heat maps, the forecast skills for Day 5 in UKMO is similar to the skills of Day 3 in NCEP and JMA models in terms of capturing the spatial pattern of local precipitation.

It is not definite that the maximum CSI will decrease with forecast days. Taking the heat maps for ECMWF model QPF as an example (Figure 4), the maximum CSI in Day 1 is about 7% lower than that of Day 2, indicating short-range model error in capturing precipitation due to uncertainties in initial condition and spin-up process. Looking into the cases captured by the two thresholds corresponding to these two forecast days, it is found that the thresholds for Day 2 capture one more heavy rain day compared with thresholds for Day 1, and the false alarm ratio (FAR; see Appendix) (Barnes *et al.* 2009) for Day 2 is about 8% lower than that of Day 1. Nevertheless, the number of days with heavy rain is quite limited and missing a heavy rain day can result in considerable drop in CSI.

5. Validation of new method in QPF for the wet season in 2018

The calibrated thresholds in the training set are applied to the verification set to test the effectiveness of forecasting light, medium and heavy rain. DMO of single-point rainfall forecasts adopted in operational reference ("QPF Summary Table") is used as baseline for comparison and their reference grid locations are listed in Table 3. Whenever more than one grid points are used in the model, the point with the highest CSI is selected in the comparison. The results of verification are plotted in Figures 8(a)-(c) for the three rainfall categories. Solid dots represent short-term forecasts (Day 1-3) and hollow dots denote medium range forecasts (Day 4-8). As the verification period is from April to September 2018, scatterplot is used where a gain in CSI in the verification period against the training period is plotted on the x-axis to investigate the effectiveness of calibration and performance during the wet season

In light rain or above category (Figure 8a), the "calibrated" QPF generally perform better (positive value in y-axis) than single point DMO. Most models show larger improvement in short range forecast than the medium range. ECMWF and UKMO benefited the most from the calibration scheme, gaining 4-12% of CSI for Day 1-3. UKMO and NCEP also have about 4% and 2% gain of CSI on average. It is not surprising to observe that all points have a positive value in the x-axis as the variability of precipitation increases in the wet season. The optimized combination of rainfall threshold and percentage of grids obtained from the training period do not necessarily guarantee similar level of the gain in CSI.

Figures 8(b)-(c) show results of medium and heavy rain respectively. It is noted that the calibrated "QPF" shows a larger gain in CSI for the heavy rain category than the case for medium rain category. The extent of improvement is clearly model dependent. In case of medium rain, the skill of precipitation forecasts from ECMWF and UKMO are generally enhanced, yet the other two models do not gain any advantage. In forecast of heavy rain, UKMO shows the largest improvement with about 30% of CSI for Day 1-3, followed by that of ECMWF at about 20% for Day 1-3. Performance of JMA model QPF can be improved slightly through calibration but there is almost no change for NCEP model QPF.

Based on the findings of spatial characteristics of model QPFs in training set and the verification results, it can be inferred that the calibrations, with whichever method, are only applicable for certain models. For models that have systematic bias in precipitation intensity while having certain skills in capturing the precipitation pattern, such as ECMWF and UKMO, the calibration would be able to improve the results from single grid points of DMO, especially for the case for heavy rain. As for those models that are not responsive towards calibration of spatial statistics such as NCEP and JMA, they are not necessarily the worse. Other means of calibration might still give meaningful results, such as extending the frequency matching method of Li *et al*. (2014) to medium-range QPF.

6. Conclusion and Future Development

As the resolution of NWP models have been increasingly fine with improved physics and data assimilation, mesoscale precipitation patterns are expected to be better resolved. The feasibility of calibrating the QPF from DMO of NWP models with their spatial statistics against the actual precipitation of Hong Kong was

studied. The proposed two calibration methods determine the optimal thresholds of criteria to maximize CSI.

This study found that the performance of ECMWF and UKMO models improved by incorporating the spatial statistics of precipitation forecasts. JMA model shows marginal improvement and NCEP is almost inert towards this algorithm.

There are certain limitations in the calibration scheme utilising the spatial cumulative distribution of models. As models are upgraded from time to time, datasets with high horizontal resolutions are not available for a long period of time. This leads to a limited sample of heavy rain cases and the calibration can only roughly capture the behaviour of models in these cases. The CSI would fluctuate a few percent for the differences of only one or two days captured in the training set under the constraints. Furthermore, in the longer forecast days, the pattern of high CSI becomes more ambiguous as models are losing forecast skills and the calibration for the precipitation categories is becoming less significant and random.

Different NWP models and calibration methods can also be blended to give a more representative post-processed QPF guidance for Hong Kong in the future. Based on the findings of this study, for instance, multi-model ensembles using different weights of the calibrated categories of ECMWF and UKMO combined with the DMO of NCEP and JMA can be tested to produce a more reliable guidance of the category of rain with a view to reducing the variance of error in QPF using the ensemble approach (Vijayakumar, 2007). For more intense precipitation such as very heavy rain (50mm or more) and above, the DMO values of the models or EPS members and blending with analogue forecast method (Chan *et al.*, 2014) would be investigated.

References

Barnes, L.R., D.M. Schultz, E.C. Gruntfest, M.H. Hayden, and C.C. Benight, 2009: CORRIGENDUM: False Alarm Rate or False Alarm Ratio?. *Wea. Forecasting*, **24**, 1452–1454, https://doi.org/10.1175/2009WAF2222300.1

Brown, J.D., and D. Seo, 2010: A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts. *J. Hydrometeor.*, **11**, 642–665, https://doi.org/10.1175/2009JHM1188.1

Chan, S.T., C.H. Fong, and M.Y. Chan, 2014: An Analog Forecast System for Precipitation Forecasting. *The Hong Kong Observatory*. https://www.hko.gov.hk/publica/reprint/r1097.pdf

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2

Donaldson, R.J., Dyer, R.M., and Kraus, M.J., 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conference on Severe Local Storms (Norman, Okla.)*, AMS, Boston, 321-326.

Duan, Q., Z. Di, J. Quan, C. Wang, W. Gong, Y. Gan, A. Ye, C. Miao, S. Miao, X. Liang, and S. Fan, 2017: Automatic Model Calibration: A New Way to Improve Numerical Weather Forecasting. *Bull. Amer. Meteor. Soc.*, **98**, 959–970, https://doi.org/10.1175/BAMS-D-15-00104.1

Japan Meteorology Agency, 2013: The Operational Numerical Weather Prediction at the Japan Meteorological Agency, Chapter 4. http://www.jma.go.jp/jma/jma-eng/jmacenter/nwp/outline2013-nwp/pdf/outline2013_04.pdf

Li, K. K., K. L. Chan, W. C. Woo, and T. L. Cheng, 2014. Improving Performances of Short-range Quantitative Precipitation Forecast through Calibration of Numerical Prediction Models (透過校準數值預報模式提升短期定量降雨預報的表現). *The Hong Kong Observatory*. https://www.hko.gov.hk/publica/reprint/r1101.pdf

Li . Z., S. P. Ballard, and D. Simonin, 2017:. Comparison of 3D-Var and 4D-Var data assimilation in an NWP-based system for precipitation nowcasting at the Met Office. *Q. J. R. Meteorol. Soc*, **144**, 404-413.

Shrestha, D.L., D.E. Robertson, J.C. Bennett, and Q.J. Wang, 2015: Improving Precipitation Forecasts by Generating Ensembles through Postprocessing. *Mon. Wea. Rev.*, **143**, 3642–3663, https://doi.org/10.1175/MWR-D-14-00329.1

Sloughter, J.M., A.E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, https://doi.org/10.1175/MWR3441.1

Vijayakumar, Sethu, 2007: The Bias–Variance Tradeoff. *University Edinburgh.* http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf

Model	Grid resolution	No. of grids within 22.0-22.6°N and 113.7-114.5°E	Forecast Days
ECMWF	0.1°	63	8
NCEP	0.25°	12	8
JMA	0.25°	12	2 (00Z run) 8 (12Z run)
UKMO	0.234°×0.156°	12	5

Table 1: NWP model data used in the study

Rainfall category	$CSI_{grids} - CSI_{mean}$ (%)		
	Day 1	Day 2	Day 3
Light	0.6609	1.2402	0.6661
Medium	1.3104	2.7760	2.5813
Heavy	3.03	5.9235	3.3481

Table 2: The difference of averaged maximum CSI over the four NWP models using the cumulative distribution method (denoted as CSI_{grids}) and the averaged maximum CSI of models using the simple mean method (denoted as CSI_{mean}), for the three thresholds and the first three forecast days.

Model	Location of grid point referenced
ECMWF	22.50°N,114.00°E
	22.25°N,114.25°E
NCEP	22.00°N,114.00°E
JMA	22.5°N,113.75°E
	22.5°N,115.00°E
UKMO	22.46°N,114.00°E
	22.31°N,114.30°E

Table 3: Grid points from NWP models to be compared against the calibrated QPF.

Figure 1: Domain in which grid points are extracted.

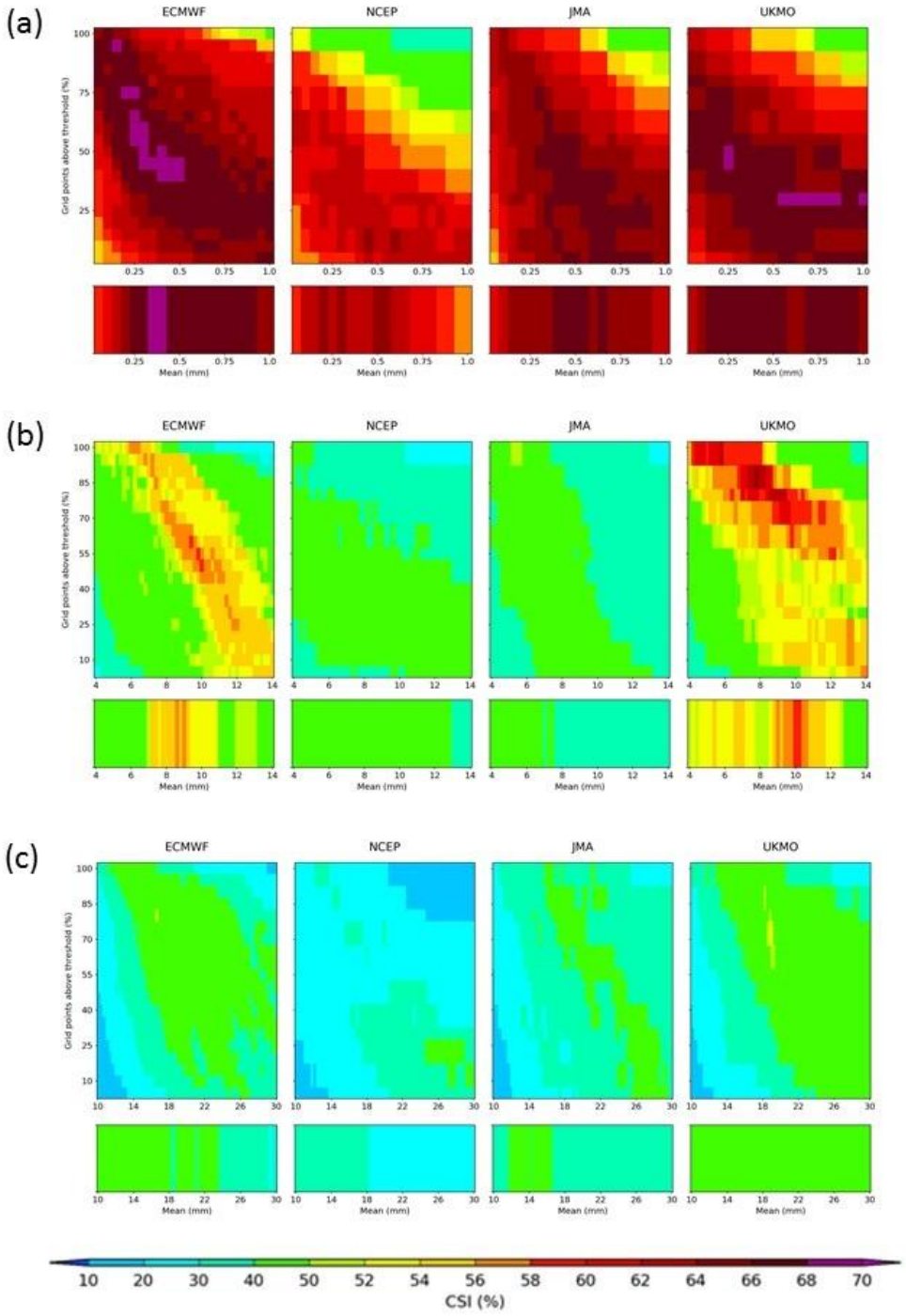

Figure 2: CSI of (a) light rain or above; (b) medium rain or above; and (c) heavy rain or above for the four NWP models of 00Z run on forecast day 2, using the cumulative distribution of intensity of precipitation (top) and simple mean (bottom)

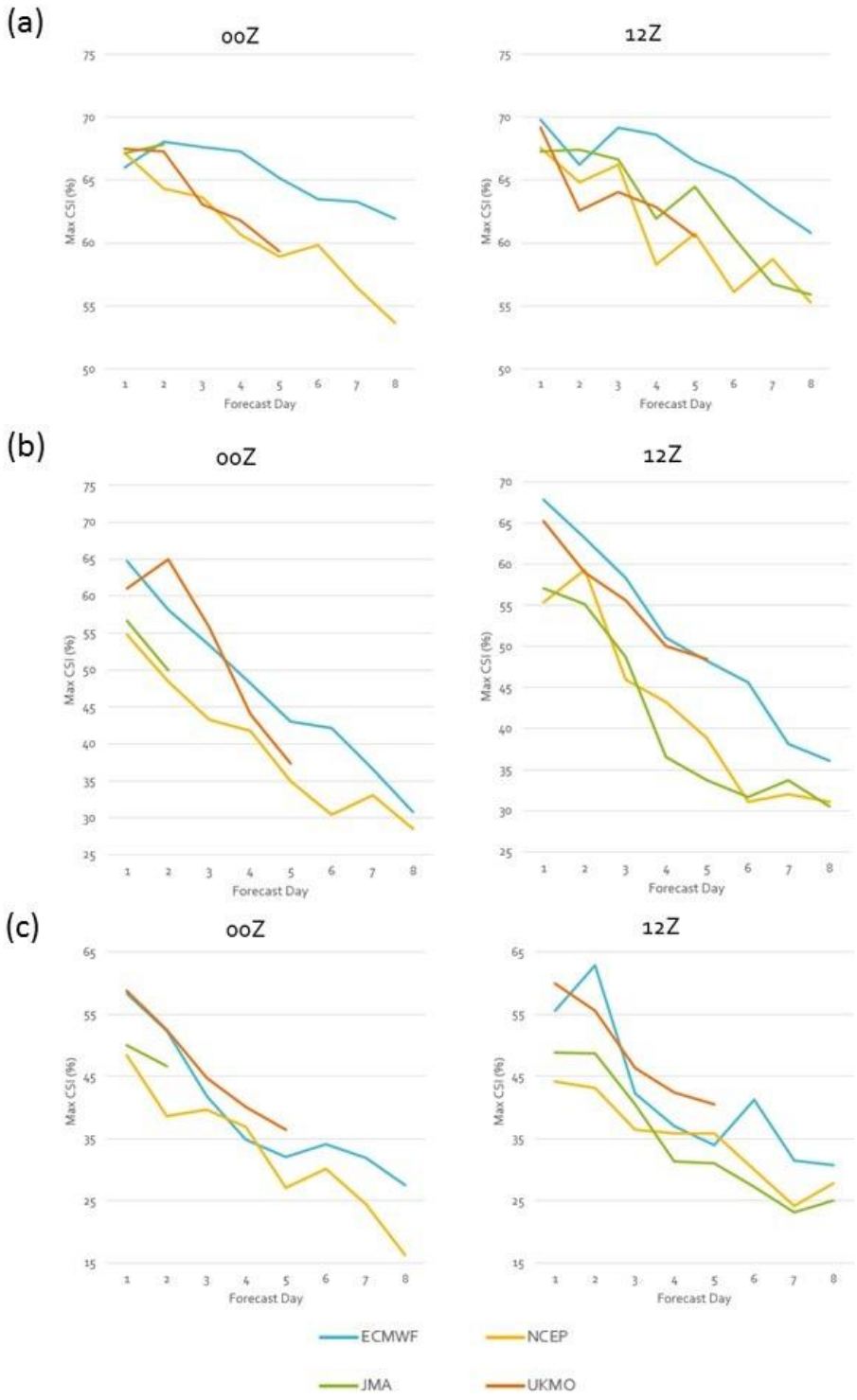

Figure 3: The CSI at the optimal threshold of each NWP model during the forecast period for (a) light rain or above; (b) medium rain or above; and (c) heavy rain or above.

Figure 4: CSI heat maps for heavy rain or above of ECMWF 12Z run. The forecast day and the threshold with the maximum CSI are included in the title of each panel.

Figure 5: Same as Figure 4 with UKMO 12Z run.

Figure 6: Same as Figure 4 with NCEP 12Z run.

Figure 7: Same as Figure 4 with JMA 12Z run.

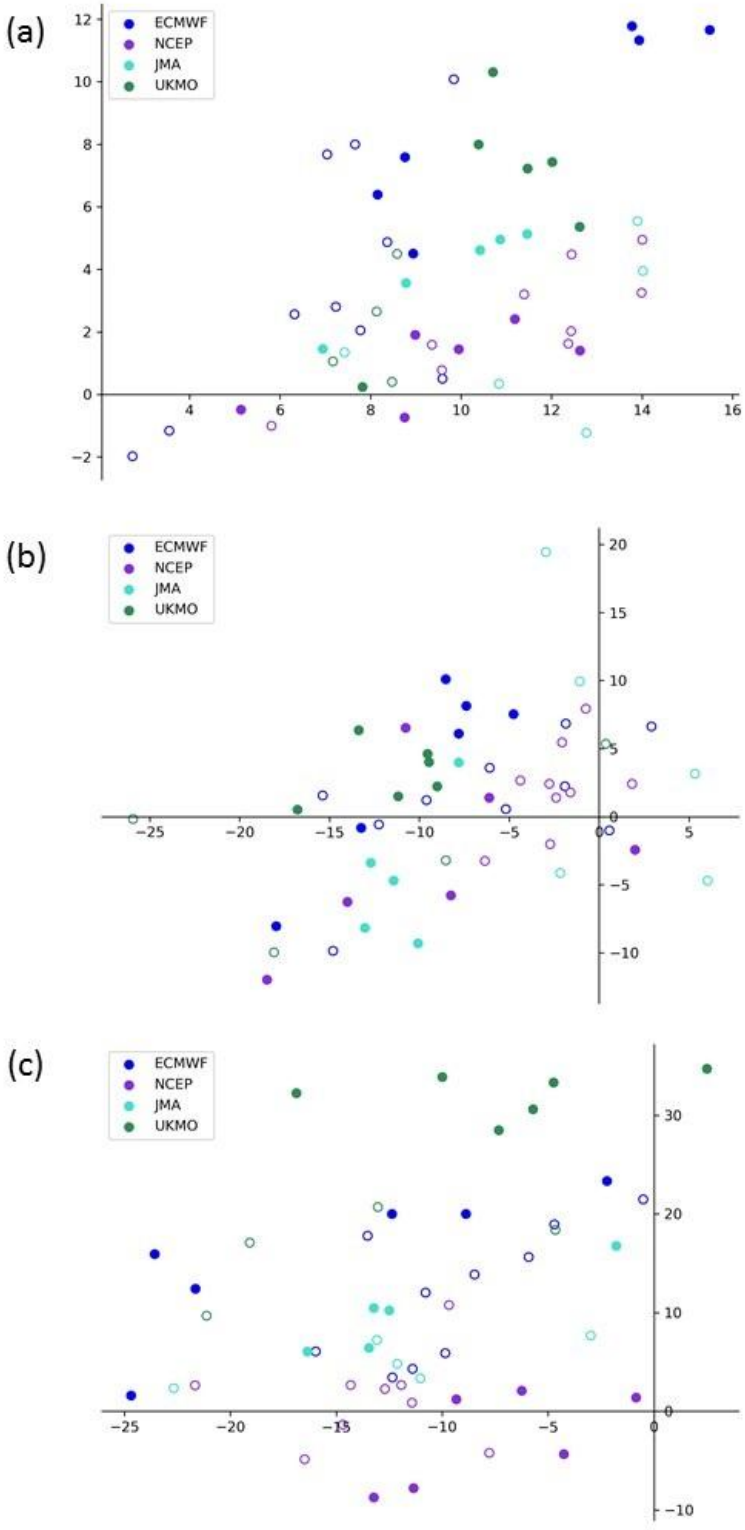

Figure 8: The performances of NWP models in the verification of (a) light rain or above; (b) medium rain or above; and (c) heavy rain or above. The y-axis is the difference of CSI of the verification set with CSI of DMO of the same period in %; The x-axis is the difference of CSI of verification set and CSI of training set under the same thresholds in %. Solid dots representing Day 1-3 forecast while hollow dots for Day4-8.

Appendix: Definition of Verification Measures

In the following contingency table of forecast and observed events,

		Event Observed	
		Yes	No
Event Forecast	Yes	*a*	*b*
	No	*c*	*d*

a, *b*, *c* and *d* are the number of events.

The verification measures are defined as follows:

Critical Success Index (CSI) = $\frac{a}{a+b+c}$, and

False Alarm Ratio (FAR) = $\frac{b}{a+b}$.