Week 3 Temperature Probabilistic Forecasts using Bayesian Processor of Ensemble

Shin-Yu Chu, Hui-Ling Chang, Shih-Chun Chou, Tsun-Wen Lo, Yun-Jing Chen
Central Weather Bureau,
Taipei, Taiwan

Abstract

A method is demonstrated in this study to calibrate the ensemble model output of the 3-week lead 2-meter weekly average temperature forecast at given weather stations in Taiwan. Bayesian Processor of Ensemble (BPE), a statistical post-processing method that utilize Bayes' Theorem, consists of procedure generating a likelihood and a prior distribution to obtain a posterior probability distribution based on latest evidence. The fully Bayesian nature allows BPE to produce informative probability distributions without relying on long training data compared to other calibration methods, allowing probabilistic forecasts based on ensemble models currently lacking long term reforecasts.

First, preprocessing of the data pair prior to likelihood generation is required to transform data to a normal-distributed form to fulfill the postulation of a meta-Gaussian model using Yeo-Johnson Transform (YJT), a method of power transform for unbounded real-valued data. Subsequently, applying a low-pass filter to remove unpredictable noise in climatological data can further improve the reliability of the BPE post-processed forecast. After pre-processing our training data, we use a Bayesian-based machine learning algorithm to infer the likelihood probability distribution. Likelihood is the marginal distribution given by the correlation between the ensemble mean and observation in the training data, and prior is given by climatological distribution of the predictand. The posterior, or the forecast probability density function for a target lead, is generated by the fusion of the inferred likelihood with the climatological prior once receiving the latest ensemble prediction. According to our validation results, ensemble model post-processed by BPE is better skilled and reliable compared to its raw counterparts in predicting week-3 mean temperature, with only few years of training data required.

Keywords: Ensemble Model, Probabilistic Forecasting, Statistical Post-processing, Bayesian Processor of Ensemble

I. Introduction

For extended range forecasts where the errors caused by non-linear process and incomplete modeling of the underlying process are amplified to a degree, deterministic forecast are no longer able to produce results that could be confidently used by decision makers. Probabilistic forecast is a preferable tool in such scenarios to quantify the uncertainty and distributions of future realizations. Using alternating perturbations and physical parameterizations, Ensemble Prediction Systems (EPS) represent the possible realizations of the current state by multi-member forecasts due to aforementioned reasons. However, EPS still inherits certain bias and are generally under-dispersive, hence, statistical post-processing is a proven approach in improving the reliability, sharpness and calibration of such models. There exits various calibration methods used by Central Weather Bureau (CWB). Each method holds its own strengths, some are simple but elegant, some are complex but sophisticated. However, some of them relies on the accumulation of large hindcast or historical templates to establish a sufficient amount of training data to effectively calibrate the EPS, some requires large amount of estimated parameters, some may suffer from over-fitting or over-weighting certain predictors. (Hodyss et al., 2016; 羅存文 等, 2016). Moreover, the raw output of EPS tends to approach the climatological distribution asymptotically as leadtime exceeds 2 weeks. In this research, we tend to confront the aforementioned issues by maximizing the usefulness of some EPS currently with limited accumulation of runs (e.g. CWB-GEPS), combining it with a longer climatological distribution by utilizing the asymptotic characteristic, sometimes to constrain predicted values to prevent over-fitting, sometimes to increase the dispersion, using concepts of Bayesian Statistics.

Bayesian Processor of Ensemble (BPE, Krzysztofowicz & Evans, 2008, hereafter KE08;Wang et al., 2018), is a novel method currently under development

in CWB. Unlike other Bayesian methods currently used to post-process EPS, the structure of BPE fully conforms to Bayes' Theorem:

$$\mathrm{P}(w_{t+l}|x_{ft}) = \frac{\mathrm{L}(x_{ft}|w_{t+l})R(w_{t+l})}{H(x_{ft})} \quad \text{(Eq. 1)}$$

Fusing a shorter training set, or likelihood distribution ($\mathrm{L}(x_{ft}|w_{t+l})$) in the context of Bayes' Theorem, with a climatological prior ($R(w_{t+l})$) to generate a calibrated posterior.

Here, x_{ft} is defined as the ensemble mean of the predictor at time t predicting leadtime l ,and w_{t+l} is defined as the realization of predictand at time $t+l$. Then posterior, $\mathrm{P}(w_{t+l}|x_{ft})$ can be interpreted as the probability of realization w at leadtime l given a forecast x_{ft}. The prominent strength of BPE are as follows:

1. The combination of a long climatic distribution to a shorter joint sample can reduce the requirement for lengthy hindcast sets, thus minimizing the resource and time to rerun after a new version of an EPS is released, allocating valuable computer resources to other more urgent matters.

2. The complete Bayesian framework ensures that the posterior generated by BPE is well calibrated and most informative, with the training data at hand. For extended leadtimes where predictability reaches the current limit of our numerical model, the informativeness of the predictor reduces to 0, the posterior PDF will auto-converge to the climatological distribution.

II. Methodology

In this study, a sample of NCEP SubX Global Ensemble Forecast Model (NCEP-SubX) 2-meter temperature (T_{2M}) downscaled using bilinear interpolation and simple height correction using moist adiabatic lapse rate to the corresponding locations of 28 Taiwan Meteorological Stations, forming a data vector $\{\mathbf{x}\}$. The downscaled T_{2M}, with the corresponding observations at target leadtime $\{\boldsymbol{w}\}$ forms a joint sample $\{\mathbf{x}, \boldsymbol{w}\}$, which will be referred as training set hereafter. Due to the non-stationarity in the time series of the predictand, and season-to-season variability in model performance, we first group the training set to cool and warm season. Where cool season consists of November, December, January, February and March, and warm season consists of May, June, July, August and September. Models are trained separately after splitting the data. Transitional seasons, defined as April and October, will be classified into warm/cool according to which has a better overall score during cross-validation. After seasonal-split, we further remove the intra-seasonal non-stationarity by standardizing the joint sample:

$$x'_k = \frac{x_{ft}-m_k}{s_k}, w'_k = \frac{w_{t+l}-m_k}{s_k} \quad \text{(Eq. 2)}$$

Where m_k and s_k is the climatological value of the predictand on day k in a year, after applying a low pass filter on the original time series $\boldsymbol{m_k}, \boldsymbol{s_k}$ to filter out high-frequency perturbations. As succeeding procedures in BPE assumes that both predictands and predictors are Gaussian, we have to ensure our data is distributed accordingly. KE08 preprocess the training data using Normal-Quantile Transform (NQT), first fitting the training data by a Weibull Distribution and inverse transformed onto a Gaussian quantile, but one of the drawbacks is which parametric distribution to use is pre-determined, reducing the flexibility of the framework. For example, precipitation where values are strictly positive with a point mass at 0, Gamma distribution should be a better choice instead of Weibull. Hence, Yeo-Johnson power transform (Yeo & Johnson, 2000), is used to transform our data pair to fit to a Gaussian Distribution in this study:

$$\psi(y,\lambda) = \begin{cases} \frac{(y-1)^\lambda}{\lambda}, & (y \geq 0, \lambda \neq 1) \\ \ln(y+1), & (y \geq 0, \lambda = 1) \\ \frac{-[(1-y)^{2-\lambda}-1]}{2-\lambda}, & (y < 0, \lambda \neq 2) \\ -\ln(y+1), & (y < 0, \lambda = 2) \end{cases} \quad \text{(Eq. 3)}$$

Where non-negative data such as precipitation, wind speed, radiation intensities etc., could use alternative power transform method such as Box-Cox transformation to fit data to a Gaussian Distribution.

Subsequently, the transformed variable are mapped onto a Normal Cumulative Distribution Function (CDF) and inversed transformed to a standard normal quantile:

$$V = Q^{-1}(G(\psi(w'_k,\lambda))),\ Z = Q^{-1}(\overline{\mathrm{K}}(\psi(x'_k,\lambda))) \quad \text{(Eq. 4)}$$

Where $(G, \overline{\mathrm{K}})$ is the Normal CDF of power transformed training pair, and Q^{-1} is the standard normal quantile. After transformation, we assume that the likelihood distribution, $\mathrm{L}(Z|V)$ follows the relation:

$$\mathrm{L}(Z|V=v) \sim N(av+b, \sigma^2) \quad \text{(Eq. 5)}$$

The mean of marginal distribution is located at $av+b$, when $V=v$, with a standard distribution of σ^2. The parameters (a, b, σ^2) are estimated in this study by a

Bayesian Machine Learning Package (Salvatier et al., 2016) using a No-U-Turn Sampler (NUTS) to infer the likelihood distribution based on the given data. For now, these parameters are derived individually for each station, which can be extended to grid points in the future.

After the likelihood parameters is estimated, forecast can be made based on the information provided by the mean of the latest ensemble run, $\overline{x_L}$ to generate a posterior distribution using following posterior parameters:

$$\mathrm{A} = \frac{a}{a^2+\sigma^2}, B = -\frac{ab}{a^2+\sigma^2}, T^2 = \frac{\sigma^2}{a^2+\sigma^2} \quad \text{(Eq. 6)}$$

The posterior CDF, $\Phi(\mathrm{w}|\overline{x_l})$ is written as:

$$\Phi(w_{t+l}|\overline{x_L}) = Q\left(\frac{1}{T}\left[Q^{-1}(G(\psi(w_k', \lambda))) - AQ^{-1}(K(\psi(\bar{x}_k', \lambda))) - B\right]\right) \quad \text{(Eq. 7)}$$

Here, the notations are shared with previous equations. Moments and probability density function (PDF) could be derived subsequently from CDF. The flowchart of the entire system is shown in Fig. 1.

III. Data

In this study, we aim to post-process ensemble output up to leadtimes of 3 weeks, with the observed weekly mean temperature as our predictor. Due to NCEP-SubX issues forecasts on Wednesday, Week 3 is defined as a leadtime of 18-24 days (432-576 hours) relative to the initial time, which starts from the third Sunday after the day the forecast is issued, similar to the definition of ECMWF(Vitart et al., 2019). The predictor is derived by the weekly ensemble mean T_{2M}, averaged from the ensemble mean of T_{2M} in leadtimes between 432-576 hours. For both warm seasons and cool seasons, three years of training data are included and each station are modeled individually using the method described in Part II. NCEP-SubX produces prediction on a weekly interval, with leadtimes up to 35 days, therefore, for each group, our training data will contain approximately $4 \times 5 \times 3 = 60$ data points. The validation set, contains data starting from the year 1999 to 2019, minus 3 years of training data, therefore, our validation data will contain approximately 17 years of data, which covers $4 \times 5 \times 17 = 340$ data points. The length of validation data ensures that our validation results describes the long-term, conclusive quality of a probabilistic forecast, including different climate regimes, not only a subset of hand-picked interval. The climatological distribution and moments are derived by the station data spanning from 1989 to 2019, if the station was established later than 1989, then the data will be taken as the longest record available. When training the model, 27 sets of parameters are trained individually to remove the data from a certain year when creating the climatology distribution. For example, if we are validating year 2001, our parameters should be trained without the climatological data from 2001, to ensure that the data from our trained model is completely independent from the validation set.

IV. Results

To fully assess the quality of a probabilistic forecast, both the benchmarks of reliability and the sharpness has to be evaluated. Hereafter, the raw ensemble will be referred as RAW and the post-processed ensemble by BPE will be referred as BPE_cal. The first metric we use to compare the RAW and BPE_cal forecast is the Continuous Ranked Probability Skill Score (CRPS, Hersbach, 2000):

$$\mathrm{CRPS} \equiv \int_{-\infty}^{\infty}[F(y) - F_o(y)]^2 dy \quad \text{(Eq. 8)}$$

$F(y)$ is the forecast probability distribution, and $F_o(y)$ is the unit step function satisfying:

$$F_o(y) = \begin{cases} 0, & x < obs \\ 1, & x \geq obs \end{cases} \quad \text{(Eq. 9)}$$

The overall CRPS then can be compared relative to the CRPS of climatological distribution, defined as Continuous Ranked Probability Skill Score (CRPSS):

$$\mathrm{CRPSS} = 1 - \frac{\overline{CRPS}_{mod}}{\overline{CRPS}_{obs}} \quad \text{(Eq. 10)}$$

The CRPSS over individual stations are shown in Fig.2 and Fig.3. For both seasons, CRPSS is improved and show skillfulness relative to climatology after calibration. We discover that for NCEP-SubX, the improvement of BPE_cal is better during cold season compared to warm season.

CRPS is a benchmark which measure both calibration and sharpness of a probability forecast, providing an overall knowledge of both the two calibration measures. For categorical, dichotomous forecast, we use the decomposed form of Brier Score to compare the reliability :

$$\mathrm{BS} = \frac{1}{n}\left[\underbrace{\sum_{i=1}^{I} N_i(y_i - \bar{o}_i)^2}_{\text{(a)}} - \underbrace{\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2}_{\text{(b)}}\right] + \underbrace{\bar{o}(1-\bar{o})}_{\text{(c)}} \quad \text{(Eq. 11)}$$

Where n denotes the total sample size, i the bin numbers, with total bin size I, and N_i the number of sample in bin i, y_i the probability value at bin i, $\bar{o}_i$ the overall relative realized frequency of bin i and $\bar{o}$ the climatological frequency. The three components of BS are (a) reliability (b) resolution and (c) uncertainty. BS is a negatively oriented score, with a lower value of reliability component and higher value of resolution component getting a better BS. Fig. 4 and Fig. 5 are the reliability diagrams and the Brier Scores of raw and post-processed ensemble. The three colored lines denotes the reliability curve when the observed value lies above, below or at normal climatological intervals. The upper and lower bounds of intervals are defined using the same method of 陳昀靖 等, 2016. A more reliable probability forecast system will produce reliability curves closer to the diagonal. From the reliability curve and the reliability component of BS, we can conclude that the reliability the BPE_cal is vastly improved over RAW.

Albeit we have shown that both reliability and resolution could be greatly improved by our method, these benchmarks could be vague for non-meteorologists. For them, the economic benefit will be the direct assessment for the quality of a probability forecast system, measured by Economic Value (Chang et al., 2015; Richardson, 2000). Due to the calculation of economic value between each users varies between their incident of interest, a benchmark for assessing the overall relative economic values between probability forecast systems for all users is used, which is defined as Informativeness Score (IS, Krzysztofowicz, 1992, hereafter K92):

$$\mathrm{IS} = \left(1 + \left(\frac{\sigma}{aS}\right)^2\right)^{-\frac{1}{2}} \quad \text{(Eq. 12)}$$

As stated in K92, a forecast system *A* that is sufficient for *B* will have an ex-ante economic value equal or higher than the latter, which in turn have an equally or higher IS. From Fig. 6.a, the IS of BPE_cal is improved with high significant level in cold season, but less so during the warm ones. Finally, we show the Calibration Score (CS, Krzysztofowicz & Sigrest, 1999), defined as:

$$\mathrm{CS} = \left\{\frac{1}{3}\left[(r_{75} - 0.75)^2 + (r_{50} - 0.5)^2 + (r_{25} - 0.25)^2\right]\right\} \quad \text{(Eq. 13)}$$

Where r_p indicates the probability of observed values lies within a designated exceedance fractile *p*, which in this case is [0.25,0.5,0.75]. CS measures how *well-calibrated* is the probability forecast, and is a negative-oriented score. From Fig. 6.b, we can see that probabilistic forecast produced by BPE_cal is more well-calibrated than its raw-counterparts conclusively, at least in the selected exceedance fractile, with a near-zero CS.

V. Conclusion and Future Developments

This is a pilot study to confirm the claim that the BPE structure allows improved skill from fusing a limited joint sample with a longer climatic sample when applied to given Taiwan stations. In our study with only 60 training data points, this seems to hold after validating a dataset of over 300 data points, which shows that BPE can improve CRPS, BS, IS and should be well-calibrated. Currently, the system is still simple, using univariate calibration. However, BPE allows combination of multiple predictors or multiple EPSs, which might produce even more robust probabilistic forecasts. In the future, we plan to:

(1) Test BPE in post-processing CWB_GEPS, currently the prominent EPS developed by CWB.
(2) Extend the structure of BPE to allow two or more predictors.
(3) Use wave-filters to select predictable wavebands, filtering unpredictable noise from the raw EPS.
(4) Adjust climatic prior to long-term climate drifting, using trend-detection technique (Chu et al., 2010).

VI. References

Chang, H.-L., Yang, S.-C., Yuan, H., Lin, P.-L., & Liou, Y.-C. (2015). Analysis of the relative operating characteristic and economic value using the LAPS ensemble prediction system in Taiwan. *Monthly Weather Review, 143*(5), 1833-1848.

Chu, P.-S., Chen, Y. R., & Schroeder, T. A. (2010). Changes in precipitation extremes in the Hawaiian Islands in a warming climate. *Journal of Climate, 23*(18), 4881-4900.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting, 15*(5), 559-570.

Hodyss, D., Satterfield, E., McLay, J., Hamill, T. M., & Scheuerer, M. (2016). Inaccuracies with multimodel postprocessing methods involving weighted, regression-corrected forecasts. *Monthly Weather Review, 144*(4), 1649-1668.

Krzysztofowicz, R. (1992). Bayesian correlation score: A utilitarian measure of forecast skill. *Monthly Weather Review, 120*(1), 208-220.

Krzysztofowicz, R., & Evans, W. B. (2008). Probabilistic Forecasts from the National Digital Forecast Database. *Weather and Forecasting, 23*(2), 270-289. doi:10.1175/2007waf2007029.1

Krzysztofowicz, R., & Sigrest, A. A. (1999). Calibration of probabilistic quantitative precipitation forecasts. *Weather and Forecasting, 14*(3), 427-442.

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.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science, 2*, e55.

Vitart, F., Balmaseda, M., Ferranti, L., Benedetti, A.,

Sarojini, B., Tietsche, S., . . . Bechtold, P. (2019). Technical Memo.

Wang, Y., Zhang, X., & Toth, Z. (2018). *Application of the Bayesian Processor of Ensemble to the Combination and Calibration of Ensemble Forecasts.* Paper presented at the International Conference On Signal And Information Processing, Networking And Computers.

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959.

陳昀靖、陳孟詩、陳重功、羅存文、王志嘉，2016: 系集動力統計預報第二週溫度三分類機率預報之開發與評比，105 年天氣分析與預報研討會，論文彙編，中央氣象局，臺灣，臺北，A3-5

羅存文、陳重功、王志嘉，2016:氣象局第 2 與 3、4 週動力統計預報作業之發展，105 年天氣分析與預報研討會，論文彙編，中央氣象局，臺灣，臺北，A6

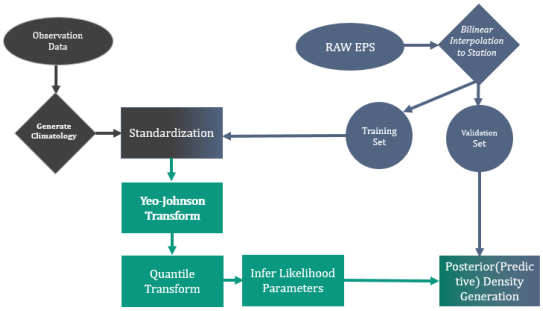


Fig 1. Flow chart of the BPE system.

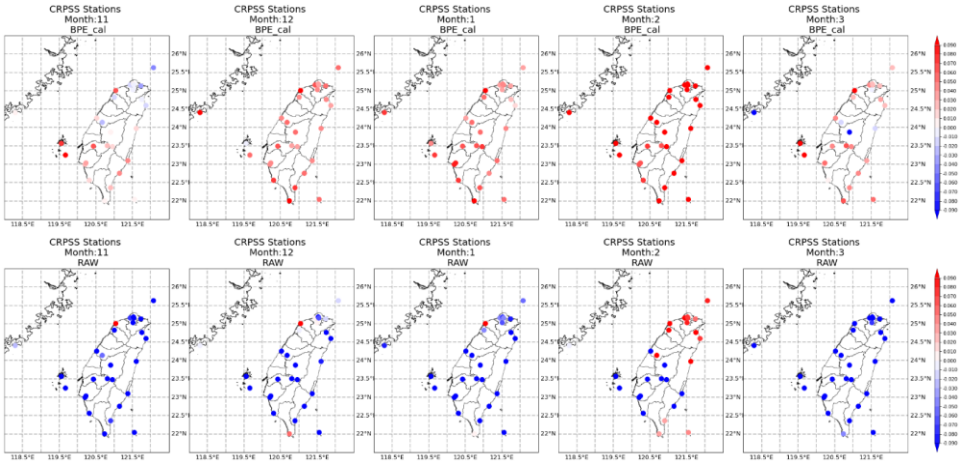


Fig.2 CRPSS of BPE_cal (upper panel) and RAW(lower panel). From left to right is Month 11 through 3 during validation period. Red corresponds to skillful predictions relative to climatology.

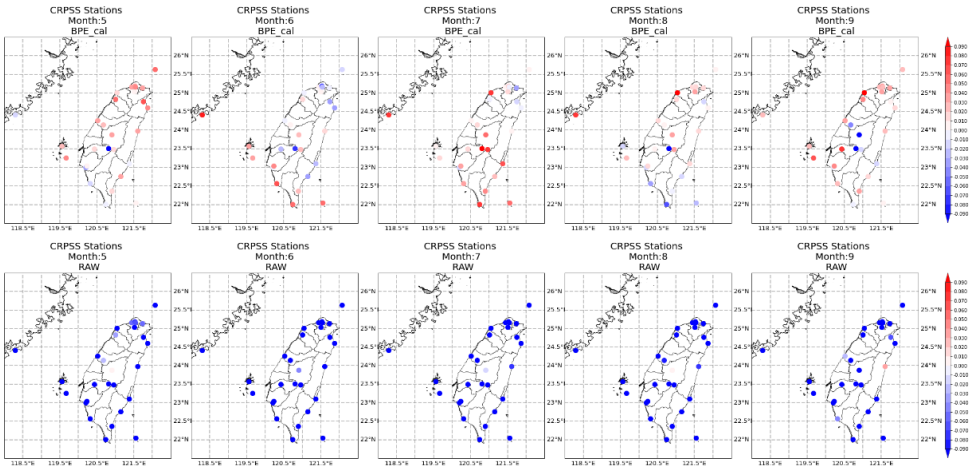
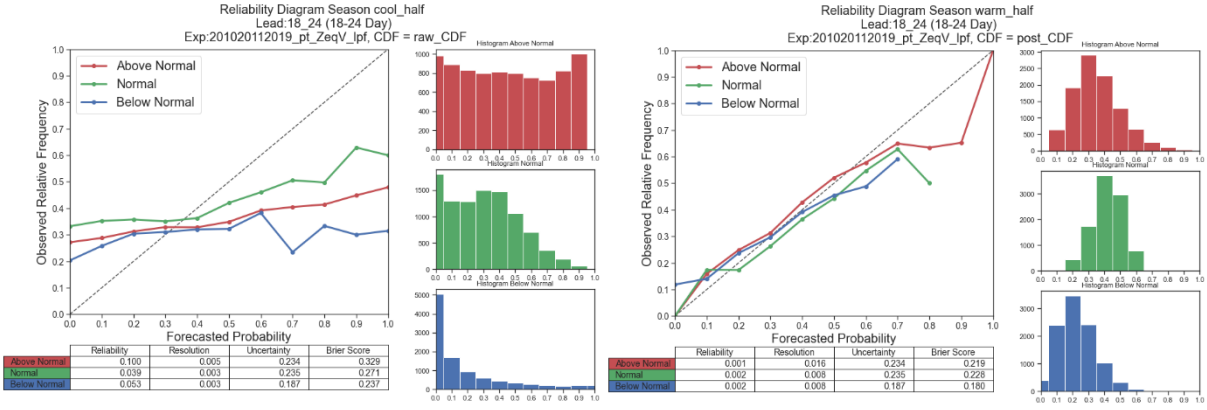


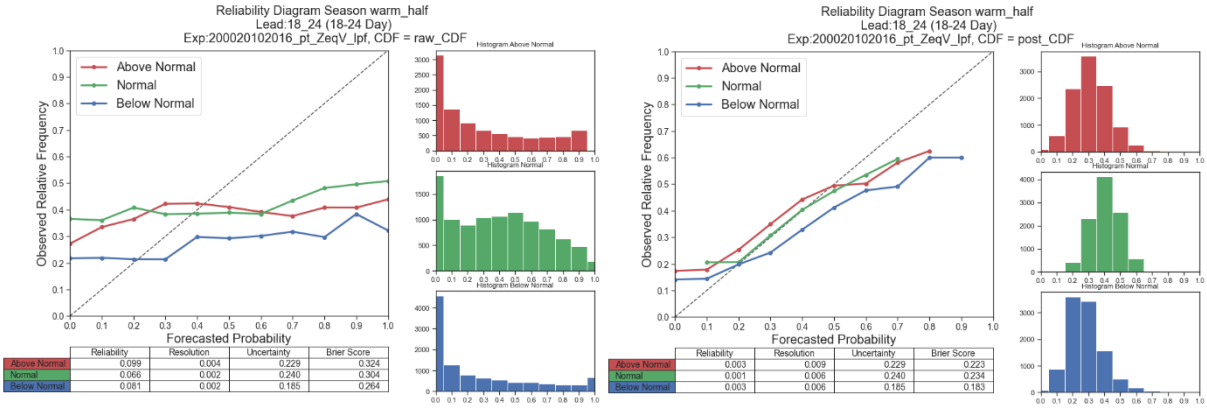
Fig. 3 Same as Fig. 2, but for month 5-9



	Reliability	Resolution	Uncertainty	Brier Score
Above Normal	0.100	0.005	0.234	0.329
Normal	0.039	0.003	0.235	0.271
Below Normal	0.053	0.003	0.187	0.237

	Reliability	Resolution	Uncertainty	Brier Score
Above Normal	0.001	0.016	0.234	0.219
Normal	0.002	0.008	0.235	0.228
Below Normal	0.002	0.008	0.187	0.180

Fig. 4 Reliability diagram of RAW (left) and BPE_cal (right) for Month 11-3 during validation period.



	Reliability	Resolution	Uncertainty	Brier Score
Above Normal	0.099	0.004	0.229	0.324
Normal	0.066	0.002	0.240	0.304
Below Normal	0.081	0.002	0.185	0.264

	Reliability	Resolution	Uncertainty	Brier Score
Above Normal	0.003	0.009	0.229	0.223
Normal	0.001	0.006	0.240	0.234
Below Normal	0.003	0.006	0.185	0.183

Fig.5 Same as Fig. 4 but for Month 5-9.

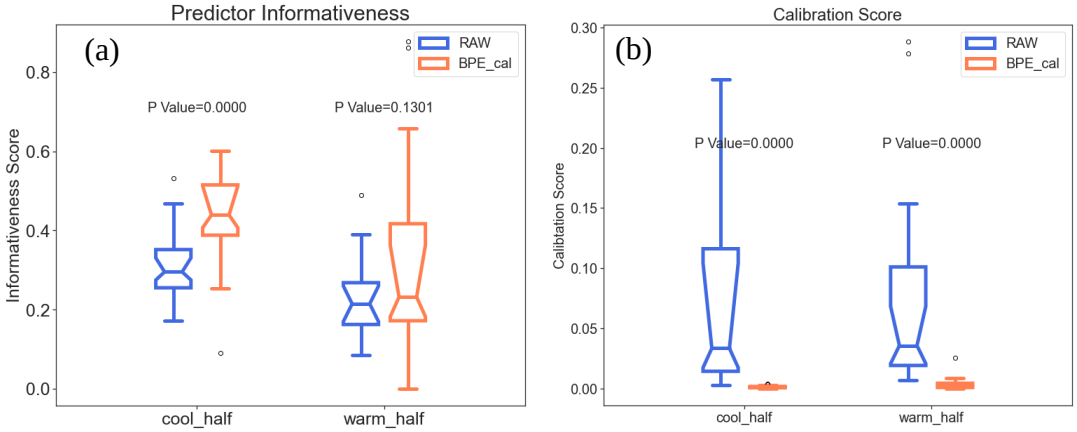


Fig. 6 Box plot of (a) The Informativeness Score (IS) and (b) the Calibration Score (CS)

貝氏系集處理器應用於第三周溫度機率預報

朱心宇 張惠玲 周柿均 羅存文 陳昀靖
中央氣象局
科技中心

關鍵字: 系集模式、機率預報、統計後處理、貝氏系集處理器