Using the Inverse of Expected Error Variance to Determine Weights of Individual Ensemble Members: Application to Temperature Prediction

Xiaogong SUN, Jinfang YIN*, and Yan ZHAO

State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081

(Received October 9, 2016; in final form November 22, 2016)

ABSTRACT

The inverse of expected error variance is utilized to determine weights of individual ensemble members based on the THORPEX (The Observing System Research and Predictability Experiment) Interactive Grand Global Ensemble (TIGGE) forecast datasets. The weights of all ensemble members are thus calculated for summer 2012, with the NCEP final operational global analysis (FNL) data as the truth. Based on the weights of all ensemble members, the variable weighted ensemble mean (VWEM) of temperature of summer 2013 is derived and compared with that from the simple equally weighted ensemble mean. The results show that VWEM has lower root-mean-square error (RMSE) as well as absolute error, and has improved the temperature prediction accuracy. The improvements are quite notable over the Tibetan Plateau and its surrounding areas; specifically, a relative improvement rate of RMSE of more than 24% in 2-m temperature is demonstrated. Moreover, the improvement rates vary slightly with the prediction lead-time (24–96 h). It is suggested that the VWEM approach be employed in operational ensemble prediction to provide guidance for weather forecasting and climate prediction.

Key words: ensemble forecast, variable weighted ensemble mean, simple equally weighted ensemble mean, prediction accuracy

Citation: Sun, X. G., J. F. Yin, and Y. Zhao, 2017: Using the inverse of expected error variance to determine weights of individual ensemble members: Application to temperature prediction. *J. Meteor. Res.*, **31**(3), 502–513, doi: 10.1007/s13351-017-6047-0.

1. Introduction

The strategy of ensemble forecasting has been broadly employed to weather prediction and climate projection to improve prediction skill. Leith (1974) pointed out that the average of a group of numerical forecasts initiated with random perturbations in initial conditions enables the provision of more accurate results than that of any single deterministic forecast. Since then, ensemble forecasting has received considerable attention and is widely utilized in operational weather forecasts. Both the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) introduced ensemble prediction for weather and climate prediction at the beginning of the 1990s (Tracton and Kalnay, 1993; Molteni et al., 1996). In addition, both global and medium-range ensemble prediction systems have been developed at various research institutions throughout the world, such as that of the NCEP (Toth and Kalnay, 1993) and the Canadian Meteorological Centre (CMC; Charron et al., 2009), The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) (Froude, 2011), the North American Ensemble Forecasting System (NAEFS) (Candille, 2009), the Institute of Atmospheric Physics Regional Ensemble Forecast System (Zhu et al., 2012), the Météo-France short-range Ensemble Prediction System (Descamps et al., 2014), and the US Navy's RELO Ensemble Prediction System (Wei et al., 2014).

Supported by the National Natural Science Foundation of China (41405006 and 91224004), Meteorological Key Technology Integration and Application Program (CMAGJ2015M85), National Key Technology Research and Development Program (2015BAK10B03), China Meteorological Administration Special Public Welfare Research Fund (GYHY201506002), and Basic Research Fund of the Chinese Academy of Meteorological Sciences (2014R016 and 2015Z003).

*Corresponding author: yinjf@camscma.cn.

Krishnamurti et al. (2009) reported the possibility of improving the threat scores and bias scores for all 10 days of forecasts from the multi-model ensemble forecasts, compared to the current best model over China. Zhu et al. (2013) pointed out that ensemble forecasts show a number of advantages for quantitative precipitation forecasts (QPFs) and probabilistic QPFs, albeit acknowledging the occurrence of systematic bias in forecasting near-surface variables. Zhou and Du (2010) suggested that ensemble-based forecasts are, in general, superior to a single control forecast, as measured both deterministically and probabilistically. Ye et al. (2014) indicated that the ECMWF ensemble prediction system's precipitation forecasts are generally skillful for flood forecasting, especially in large river sub-basins. In short, ensemble-based forecasts are generally superior to a single control forecast.

Previous studies have confirmed that ensemble forecasting has some potential benefits for weather and climate predictions (Bauer et al., 2015). It is important to quickly extract as much useful information as possible from the vast amount of ensemble forecast data. Accordingly, numerous attempts have been undertaken to generate different ensemble products that can be deployed effectively in improving the forecast skill (e.g., Tebaldi and Knutti, 2007; Leutbecher and Lang, 2014; Smith et al., 2014). Among these approaches, the simplest method is to average all of the ensemble member forecasts by assuming an equal weighting of each member. However, although the simple equally weighted ensemble mean (SEWEM) method can sometimes substantially improve the forecast skill (Du, 2007; Qi et al., 2014), the performance of individual members is certainly not equal for every single predictive event. In view of the shortcomings of the SEWEM method, much attention has been devoted to developing novel methods for a variable weighted ensemble mean (VWEM). Raftery et al. (2005) proposed a statistical method for post-processing ensembles based on Bayesian model averaging (BMA), which has been widely utilized to combine predictive distributions from different sources. For example, Liu and Xie (2014) introduced BMA to improve QPFs over the Huaihe basin of China. Du and Zhou (2011) put forward a dynamical performance-ranking method to predict the relative performance of individual ensemble members by assuming that the ensemble mean is a good estimation of the truth, which has several advantages for improving forecast skill. Krishnamurti et al. (2000a, b, 2009) established a useful algorithm based on multiple regression of multi-model solutions towards observed fields during a training period, which shows promise for simulations of seasonal climate, global weather, and hurricane track and intensity. Jewson (2013) employed a method for eliminating double counting in multi-model ensemble forecasts, which derives weights from empirically estimated correlations between the outputs from the ensemble members in the ensemble, without reference to observations. Van Schaeybroeck and Vannitsem (2014) developed a linear post-processing approach, which provides a considerable improvement in skill as compared to the traditional ensemble mean. Zhi et al. (2012) compared three kinds of multi-model ensemble forecast techniques based on TIGGE data, and pointed out that each technique has its own advantages and disadvantages. We concur that there are definite advantages to each approach thus far developed and reported in the literature—all of them are able in certain circumstances or under certain conditions to provide improved results. Consequently, such approaches can be used to improve the accuracy of numerical forecasts according to the requirements of end-users (Weigel et al., 2008).

Xie and Arkin (1996) first used the inverse of expected error variance to combine precipitation data from gauge observations, satellite estimates, and numerical model results. Almost at the same time, Huffman et al. (1995) developed a similar algorithm to compute global gridded fields of monthly precipitation, and Huffman et al. (1997) produced weighted average QPFs by weighting each model QPF according to the inverse of expected error variance over the prior two-month period. Although there are some differences between the two algorithms, both take the inverse of expected error variance into account for weights of different data sources. Ebert (2001) reported that the weighted average QPF produced by the inverse of expected error variance performed almost the same as the ensemble mean of other methods. Additionally, the inverse of expected error variance might yield unreasonable weights of ensemble members because the precipitation is discontinuous in spatial distribution although the inverse of expected error variance has been extensively applied to combine precipitation from different observations and/or model outputs.

In this study, the inverse of expected error variance is employed to produce different weights for temperature prediction by different ensemble members. The weights of ensemble members are derived based on the differences between model outputs and observations. Members that have large biases in the ensemble will be down-weighted, while ensemble members with low biases will be up-weighted. The specific procedure for determining the weights is described in detail in Subsection 2.2. It

should be noted that the weight of each ensemble member varies spatially—it is not a constant in all grids and at all levels.

Following this introduction, Section 2 introduces the technical details of the proposed method. The results of applying the method to an ensemble mean for temperature prediction is presented in Section 3. Finally, a summary and discussion are given in Section 4.

2. Data and method

2.1 *Data*

The current paper exploits the TIGGE control run datasets of the THORPEX program, provided by the ECWMF, for the summers of 2012 and 2013, and the NCEP final operational global analysis (FNL) data from global data assimilation system (GDAS). The models included in this study are those of the CMA (China Meteorological Administration), ECMWF (Europe), NCEP (USA), CPTEC (Centro de Previsão de Tempo e Estudos Climáticos, Brazil), KMA (Korean Meteorological Administration), CMC and UKMO (UK Met Office). The TIGGE data, temperature at both 2 m above the ground and standard pressure levels (1000, 925, 850, 700, 500, 300, 250, 200 hPa), over a 96-h prediction starting at 0000 UTC, are collected from the ECMWF server. All of the data are interpolated into 1° × 1° grids, which have the same horizontal grid spacing as the FNL data. The FNL data are taken as the truth to determine the performance of the TIGGE datasets. It should be noted that the model data of BoM (Bureau of Meteorology, Australia), JMA (Japan Meteorological Agency, Japan), and Météo-France (France) are not employed due to their large amounts of missing data. For further details of the TIGGE program and data, please see Bougeault et al. (2010) and Froude (2011). TIGGE data have been widely used to investigate precipitation forecasting (e.g, Zhao et al., 2010; Wang and Zhi, 2015), temperature prediction (e.g, Lin et al., 2009; Cui and Zhi, 2013; Zhi et al., 2013), and tropical cyclone tracks and intensity (e.g., He et al., 2015; Zhang et al., 2015). It should be noted that, although the TIGGE forecasts have lead-times of up to two weeks, we focus our attention on the short-term (first four days) predictions in this study.

2.2 *Method*

2.2.1 *VWEM*

Numerous studies have reported that there are differences, sometimes even considerable differences, among outcomes derived from ensemble systems, due to the differences in initial conditions or physics parameterization schemes used in an ensemble. However, it is still hard to judge which member is better (less bias). Here, the algorithm of Xie and Arkin (1996) is introduced to determine a different weight for each individual ensemble member, and the principle of the algorithm is described as follows. We consider a set of point prediction errors (Δ_i) from a set of m models, and the error between observation and prediction at any given time and location is then expressed as

$$\Delta_i(x,y,z,t) = X(x,y,z,t) - L_i(x,y,z,t),\ i \in [1,\mathrm{m}], \quad (1)$$

where X is observations, and L_i is predictions from the ith model output of an ensemble. Based on a series of errors from a set of n history files of each ensemble member, the "variance estimate" σ_i^2 of the predictions can be achieved via

$$\sigma_i^2(x,y,z) = \frac{\sum_{t=1}^{n} \Delta_i(x,y,z,t)\,\Delta_i(x,y,z,t)}{n}. \quad (2)$$

Assuming the Δ_i is random, unbiased, and normally distributed, and the errors from different models are independent, the weight (p_i) of the ensemble member i is then defined as

$$p_i(x,y,z) = \frac{\sigma_0^2}{\sigma_i^2(x,y,z)}. \quad (3)$$

Here, σ_0^2 is a unit "weight variance", which can be of any value, regardless of the value of σ_i^2. The value of $\sigma_0^2 = 1$ is chosen here.

The best estimate for prediction L_i at $(t + 1)$ is defined as $\hat{X}$, and the adjusted amount V_i is given by

$$V_i(x,y,z,t+1) = \hat{X}(x,y,z,t+1) - L_i(x,y,z,t+1),\ i \in [1,\mathrm{m}]. \quad (4)$$

In matrix form, Eq. (4) can be written as

$$\underset{m\times 1}{\boldsymbol{V}} = \underset{m\times 1}{\boldsymbol{A}}\,\underset{m\times 1}{\hat{\boldsymbol{X}}} - \underset{m\times 1}{\boldsymbol{L}}, \quad (5)$$

where $\underset{m\times 1}{\boldsymbol{V}} = [V_1, V_2, \ldots, V_m]^{\mathrm{T}}$, $\underset{m\times 1}{\boldsymbol{A}} = [1, 1, \ldots, 1]^{\mathrm{T}}$, and $\underset{m\times 1}{\boldsymbol{L}} = [L_1, L_2, \ldots, L_m]^{\mathrm{T}}$.

In order to solve $\hat{\boldsymbol{X}}$, the task is to minimize the $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{V}$ according to the principle of the least-squares method. This then gives

$$\hat{X}(x,y,z,t+1) = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{L} = \frac{\sum_{i=1}^{m} p_i(x,y,z)\,L_i(x,y,z,t+1)}{\sum_{i=1}^{m} p_i(x,y,z)}. \quad (6)$$

Here,

$$\boldsymbol{P} = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & p_m \end{bmatrix},$$

is the diagonal matrix of weights. The best estimate $\hat{X}$ in Eq. (6) is closely linked to the weights of ensemble members, and $\hat{X}$ is named as the VWEM in this work.

2.2.2 *SEWEM*

The SEWEM at $(t + 1)$ is calculated with the same weights of all members, which is simply given by

$$\mathrm{SEWEM}\,(x,y,z,t+1) = \frac{1}{m}\sum_{i=1}^{m} L_i\,(x,y,z,t+1)\,, \quad i \in [1,\mathrm{m}]\,. \tag{7}$$

2.2.3 *Absolute error and root-mean square error*

Absolute error (AE) and root-mean-square error (RMSE) are widely utilized to judge model performance. The AE is given by

$$\mathrm{AE} = |L_i\,(x,y,z) - X\,(x,y,z)|\,, \tag{8}$$

and the RMSE can be expressed as

$$\mathrm{RMSE} = \left[\frac{1}{N}\sum_{j=1}^{N}\left(L_j\,(\mathrm{x},\mathrm{y},\mathrm{z}) - X\,(x,y,z)\right)^2\right]^{\frac{1}{2}}. \tag{9}$$

Here, N is the total grid number.

The accuracy of an ensemble mean will be judged in terms of the RMSE via the VWEM and SEWEM methods. We use RMSEn and RMSEo to represent the RMSE for the VWEM and SEWEM, respectively. The fractional percentage improvement (I) of the RMSE can be defined as follows:

$$I = \frac{\mathrm{RMSEo} - \mathrm{RMSEn}}{\mathrm{RMSEo}} \times 100\%. \tag{10}$$

The VWEM method is applied to the TIGGE data within a domain over the Northern Hemisphere for validation, as illustrated in Fig. 1. The datasets at both surface level and standard pressure levels in summer (1 June to 31 August) of 2012 are utilized to calculate the weight (P_{Li}) of each ensemble member according to Eq. (3). Then, the ensemble means for summer (1 June to 31 August) of 2013 are given based on the Eq. (6), in which the P_{Li} values are calculated from Eq. (3), with the TIGGE and FNL data, during the period of summer 2012. Finally, the results from the VWEM method are compared to those calculated by the SEWEM method.

3. Results

3.1 *Performance of ensemble members*

The distributions of individual ensemble member weights for the 24-h temperature forecast at the 2-m level are shown in Fig. 2. According to the definition of the weight, the larger the value of the weight is, the better the ensemble member performs. As can be seen from Fig. 2, the CMA, UKMO, KMA, and NCEP models perform well over the Pacific Ocean, eastern Atlantic Ocean, Arabian Sea, Bay of Bengal, and South China Sea regions. The CPTEC model has a small weight value, indicating a large bias of the CPTEC model prediction. Figure 2 also demonstrates that the models perform better over ocean than over land. It should be noted that taking the NCEP-FNL data as the truth may influence the weight of individual ensemble members. For example, the ECMWF model has a lower weight compared with that of the NCEP forecasts. However, it has been reported in previous studies that the ECMWF's forecasts are normally better than those of other models (e.g., Cui et al., 2000; Wang and Zeng, 2012; Magnusson et al., 2014). The distributions of the weights at other predictive hours, that is, 48, 72, and 96 h (figure omitted), are similar to those for 24 h.

The weight distribution of individual ensemble members for the 24-h temperature prediction at 500 hPa are shown in Fig. 3. The UKMO, KMA, and NCEP models show good results over the Pacific Ocean and continental

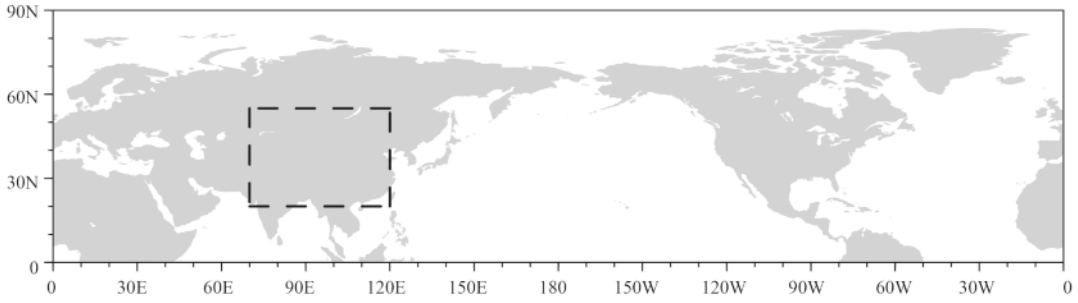

Fig. 1. Geographical coverage of the validation domain. The gray shaded areas represent land, and the Tibetan Plateau and its surrounding areas (20.0°–55.0°N, 75.0°–120.0°E) are indicated by the dashed box.

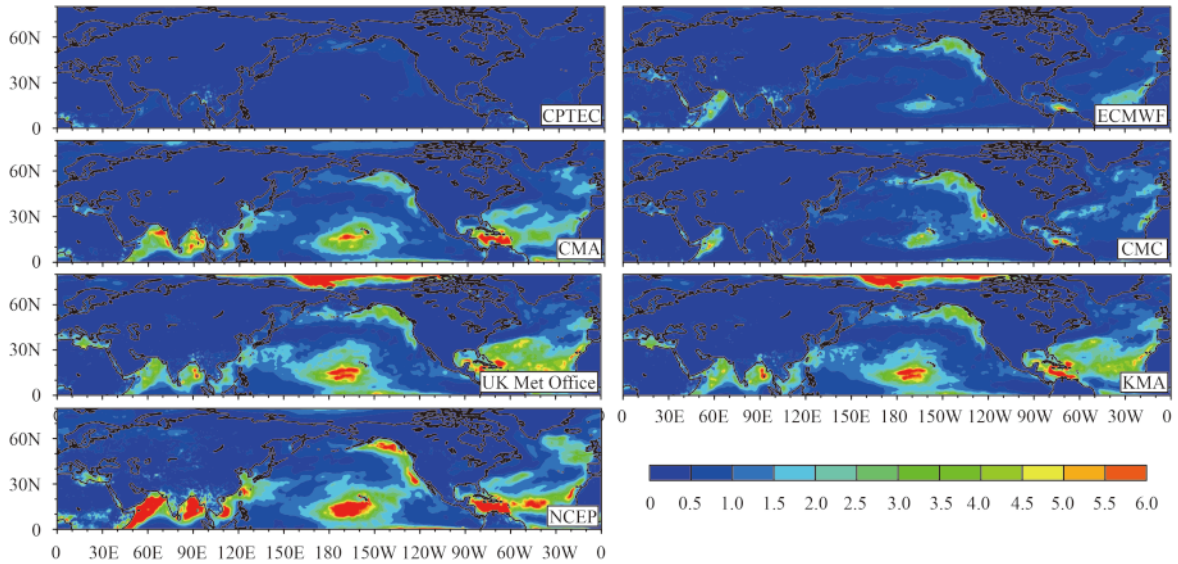

Fig. 2. Weight distributions of individual ensemble members for the 24-h temperature prediction at 2 m above the ground. The weights are calculated from the TIGGE data, with NCEP-FNL as the truth, during the period from 1 June to 31 August 2012.

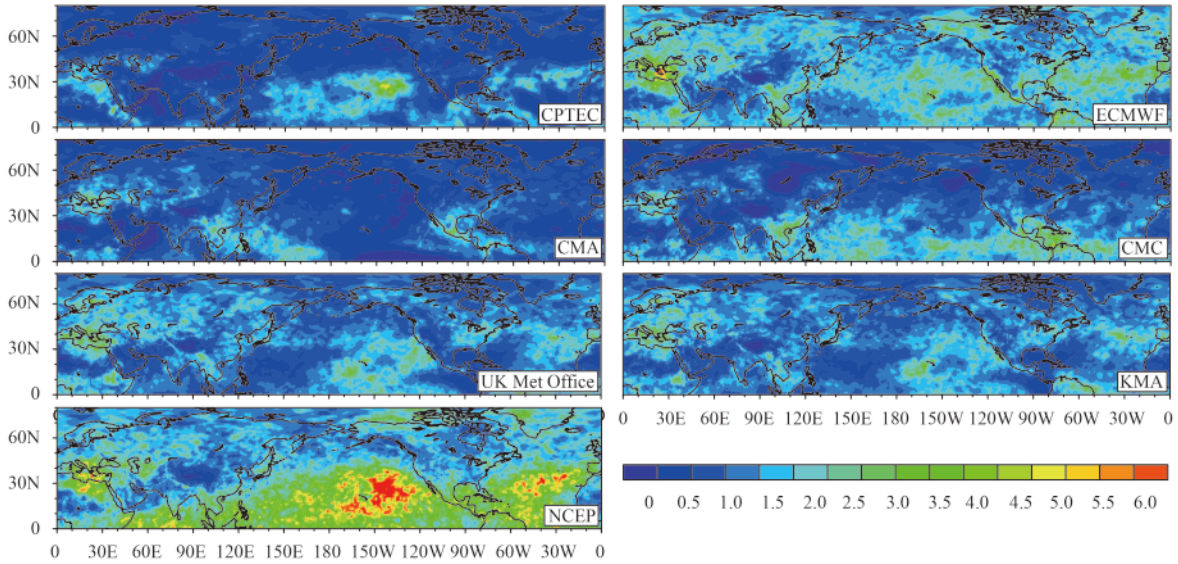

Fig. 3. As in Fig. 2, but for 500 hPa.

North America. The ECMWF and CMA models have good prediction over South China. The UKMO and CMC models provide a large weight contribution in Europe. Comparing the weights among models finds that all models appear to produce better prediction in Europe, North America, and the Pacific regions; however, they are unable to provide good results over the Tibetan Plateau and in the areas north of 50°N. Similar to the weight distribution for 2-m temperature, the models show better prediction over ocean than over land. The weight distributions at other prediction lead-times (e.g., 48, 72, and 96 h; figure omitted), are similar to those at 24 h.

3.2 *Forecast biases*

3.2.1 *Temperature at 2 m*

Figure 4 shows the averaged AE of 2-m temperature at the prediction lead-time of 24 h, from both the SEWEM and VWEM methods. As can be seen, there is a smaller AE for the ensemble mean calculated by the VWEM method, as compared with that calculated by the SEWEM method. Generally speaking, the VWEM method reduces the AE significantly, especially over the Tibetan Plateau and its surrounding areas. Both the SEWEM and VWEM method show small AE over most of Europe, western North America, northern Pacific, and northern

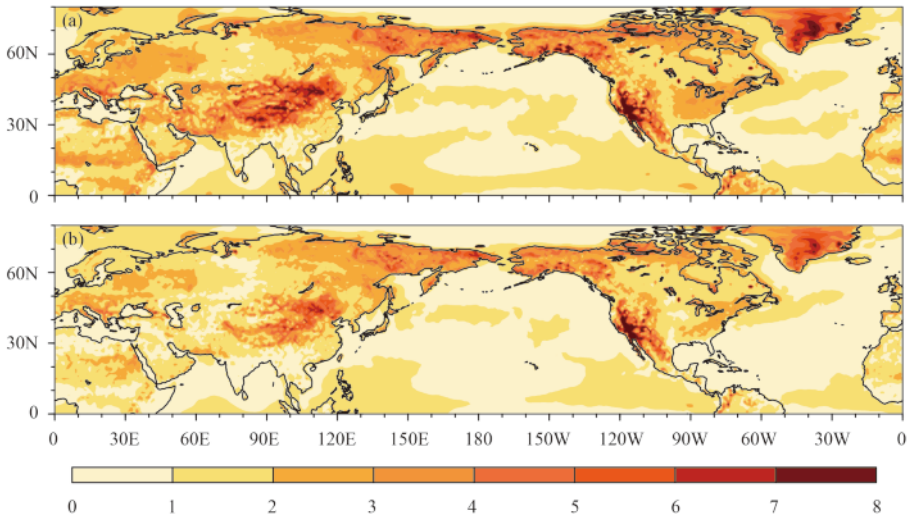

Fig. 4. Averaged AE (absolute error) of the 24-h temperature forecast at 2 m above ground. (a) Simple equally weighted ensemble mean method and (b) variable weighted ensemble mean method.

Atlantic. Forecast biases are particularly noticeable over North China, west coast of Mexico, and Greenland. Again, the results are similar to those at other prediction lead-times, that is, 48, 72, and 96 h (figure omitted).

Figure 5 shows the differences in the averaged AE between the SEWEM and VWEM methods (SEWEM minus VWEM) for 2-m temperature at the prediction lead-times of 24, 48, 72, and 96 h. Positive values are universally apparent, except in a few areas, which indicates that the VWEM method offers a significant improvement in forecast accuracy over most of the chosen domain (Fig .1). In particular, the improvement is very noticeable over the Tibetan Plateau and its surrounding areas. However, the VWEM method brings a negative effect on 2-m temperature prediction over the Sahara Desert and eastern Baikal Lake regions. Thus, further attention should be paid to the VWEM method over those regions in the future. As mentioned above, the models perform well over the oceans. Consequently, the improvement is greater over land than over ocean. It should be noted, however, that there is still improvement over ocean areas. Finally, as can be seen, there is a slight difference among the improvement level at different lead-times.

The SEWEM method has averaged RMSEs of 2.5, 2.5, 2.6, and 2.7°C at the lead-times of 24, 48, 72, and 96 h, respectively (Fig. 6). Meanwhile, the VWEM method has lower averaged RMSEs of 2.1, 2.2, 2.2, and 2.3°C, respectively. Although the actual RMSE values are still very large, the improvement is noticeable. The improvement percentages of the RMSE are 15.5%, 14.3%, 13.3%, and 12.1% at the lead-times of 24, 48, 72, and 96 h, respectively. It is apparent that the improvement is approximately 0.4°C at all lead-times, indicating that the absolute improvement is almost constant with an increase in the lead-time. It should be noted, however, that the improvement percentage of RMSE decreases with an increase in lead-time, due to the increase in absolute RMSE.

The above results indicate a significant improvement over the Tibetan Plateau and its surrounding areas (dashed box in Fig. 1). Next, we provide a detailed analysis of the RMSE over the Tibetan Plateau and its surrounding areas (Fig. 7). According to the statistical results, the improvement percentage of the RMSE is almost 25.0%, and the absolute improvement is about 1.0°C. It should be noted, however, that the models have a larger RMSE over the Tibetan Plateau and its surrounding areas than over the other areas in the Northern Hemisphere. This is due to the fact that the vast and complex topography of the Tibetan Plateau has a significant influence in numerical model simulations, as reported in numerous studies (e.g., Duan et al., 2012; Liu and Dong, 2013; Chen and Bordoni, 2014; Chen et al., 2014).

3.2.2 *Temperature at pressure levels*

Figure 8 shows the averaged AE for the 24-h temperature forecast at different pressure levels (1000, 925, 850, 700, 500, 300, 250, and 200 hPa). Generally, the AE of VWEM is far less than that of SEWEM. This indicates that the VWEM method provides better results than the SEWEM method. As for spatial distribution, large AE is apparent over the regions to the north of 30°N, while the ensemble mean shows a small AE over tropical regions, especially at 300 hPa and upper levels. Comparing the AE at different forecast levels, it can be seen that the AE is smallest at 500 hPa and largest at 1000 hPa. This sug-

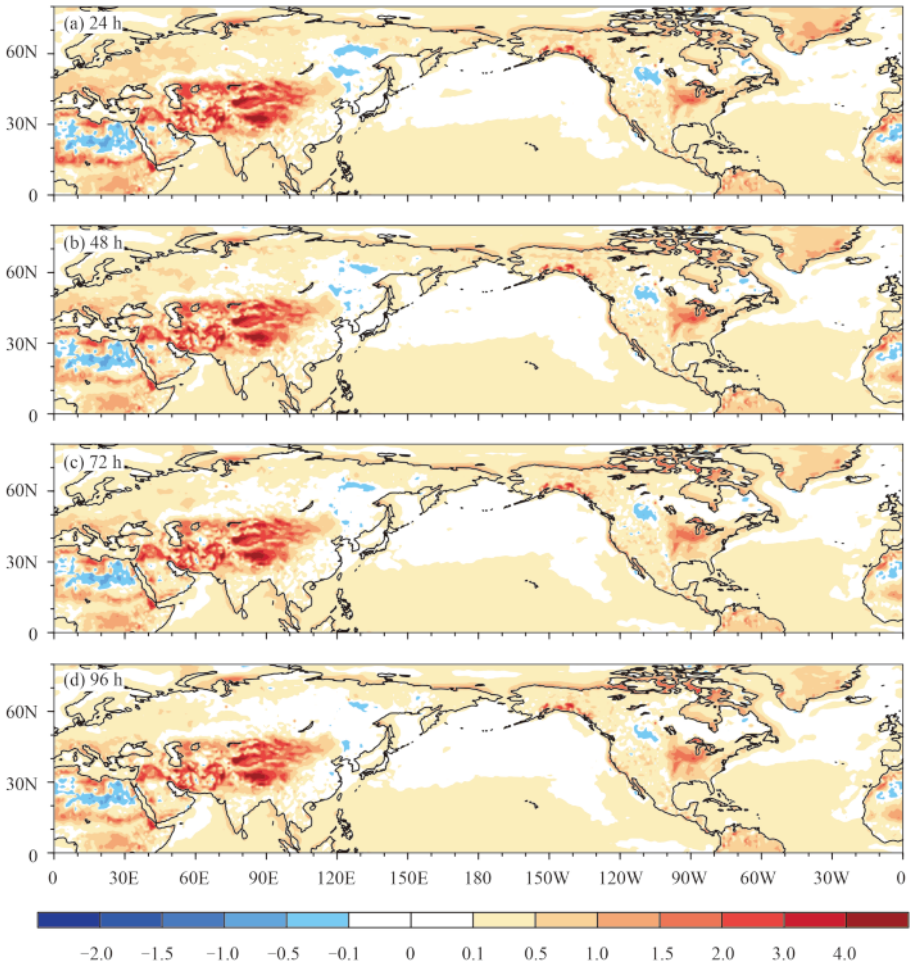

Fig. 5. Difference in averaged absolute error between the simple equally weighted ensemble mean method and variable weighted ensemble mean method for 2-m temperature, at prediction lead-times of (a) 24, (b) 48, (c) 72, and (d) 96 h.

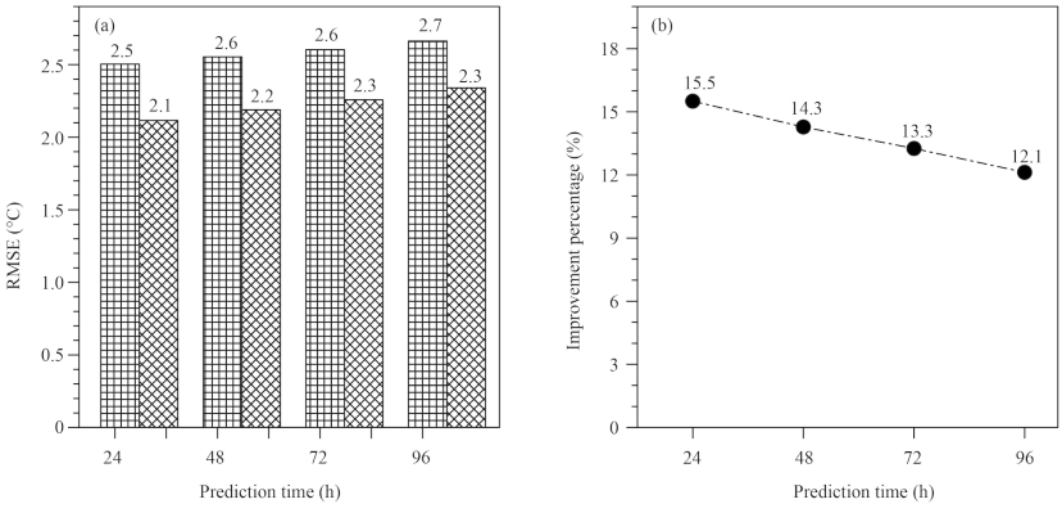

Fig. 6. The (a) averaged root-mean-square error (RMSE) of ensemble means over the Northern Hemisphere, using the simple equally weighted ensemble mean method (grid-patterned bars) and variable weighted ensemble mean method (diagonal striped bars), for 2-m temperature, at prediction lead-times of 24, 48, 72, and 96 h, and (b) the corresponding improvement percentages of the RMSE.

gests that the TIGGE ensemble predictions are more accurate in the middle troposphere than in the lower and upper troposphere. This in turn suggests that land–ocean surface processes have a strong influence on model res-

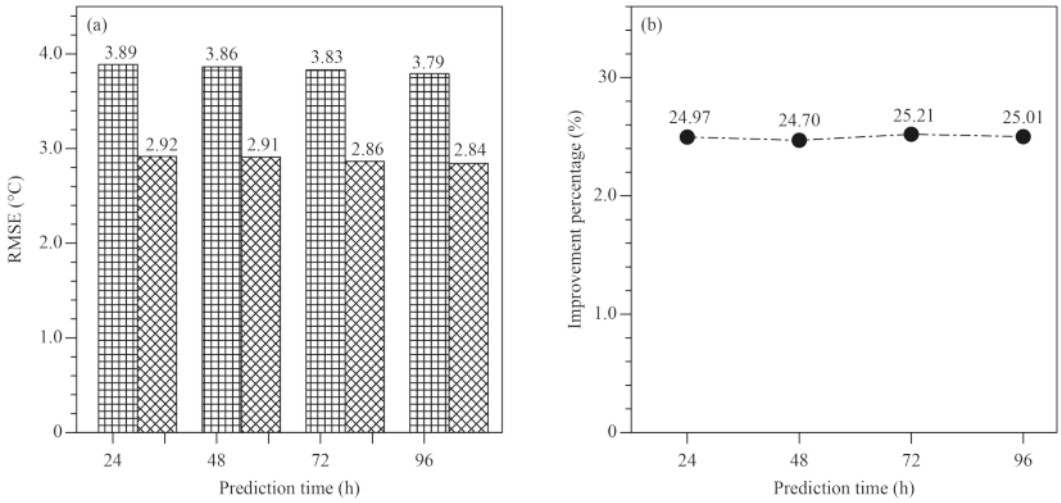

Fig. 7. As in Fig. 6, but for the Tibetan Plateau and its surrounding areas indicated by the dashed box in Fig. 1.

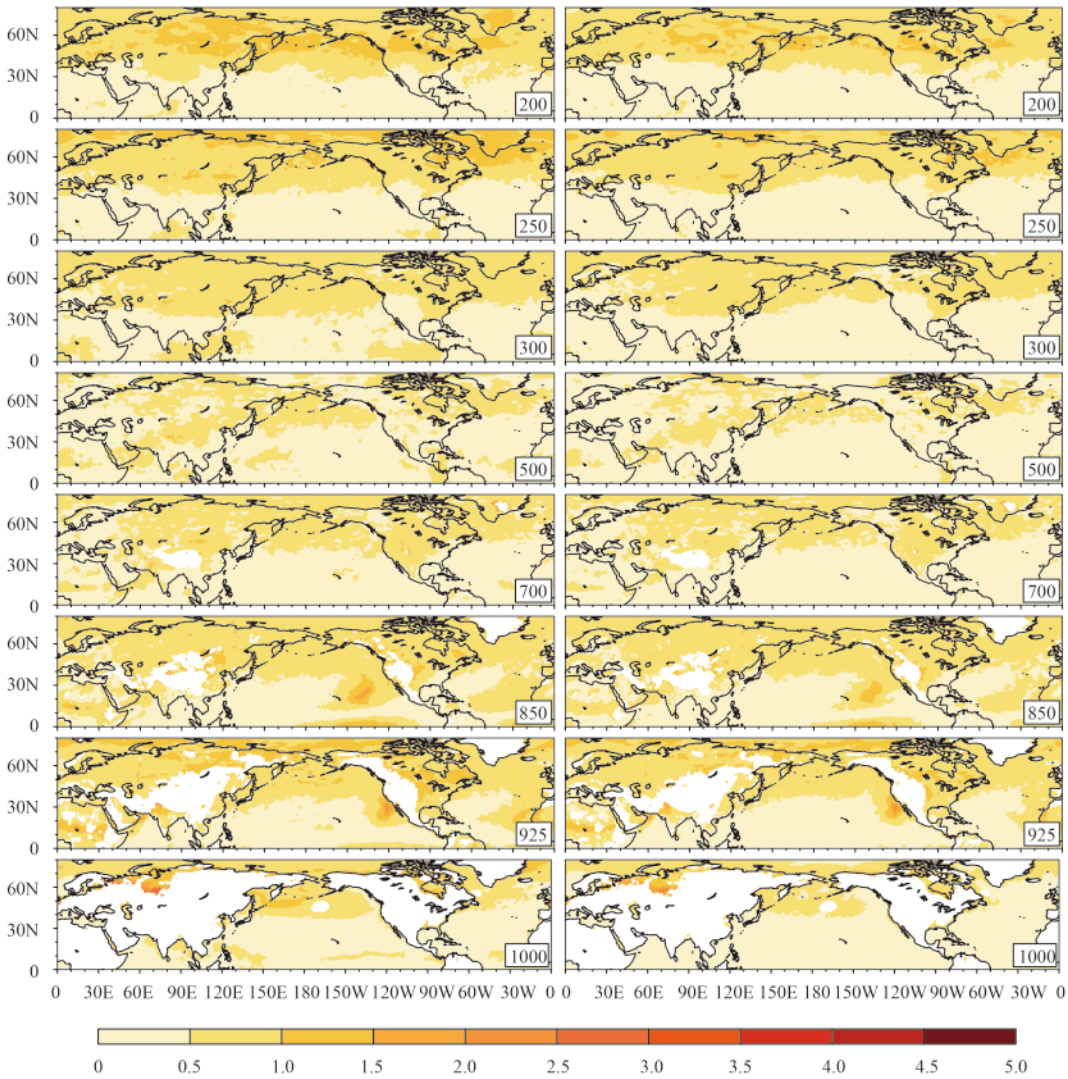

Fig. 8. Averaged absolute error at various pressure levels ranging from 200 to 1000 hPa (indicated in the lower-right corner of each panel), at the prediction lead-time of 24 h, by using the simple equally weighted ensemble mean method (left-hand panels) and the variable weighted ensemble mean method (right-hand panels). White-colored areas represent the height of terrain beyond that of the pressure level.

ults. It should be noted that the patterns of averaged AE for the lead-times of 48 and 72 h are similar to that of 24 h, except for an increase in AE.

The distributions of averaged AE for the 96-h temperature forecast at different pressure levels are shown in Fig. 9. Similar to the 24-h forecast (Fig. 8), it is clear that the AE of VWEM is less than that of SEWEM. It can also be concluded that the VWEM method provides a better result than that of the SEWEM method for the 96-h temperature forecast. However, the averaged AE for the 96-h temperature forecast is larger than that of the 24-h forecast.

The RMSEs of ensemble mean for temperature forecast at different pressure levels and at lead-times of 24, 48, 72, and 96 h, are shown in Fig. 10, separately calculated by using the SEWEM and VWEM methods. There are smaller RMSEs for the ensemble mean calculated by the VWEM method, as compared to those calculated by the SEWEM method. This indicates that the VWEM method improves the forecasting accuracy significantly. The improvements at lower levels are superior to those at upper levels. One of the reasons for this is that the prediction accuracy at upper levels is higher than that at lower levels. It should be emphasized that the VWEM method reduces the AE by almost the same amount with an increase in lead-time at the same level. For example, at 850 hPa, the RMSE is reduced by 0.06, 0.05, 0.05, and 0.05°C at the lead-time of 24, 48, 72, and 96 h, respectively. However, the RMSE increases with an increase in prediction lead-time. Consequently, the improvement percentages show a decreasing tendency with increasing lead-time. In summary, it can be concluded that the VWEM method provides a clear improvement in the forecasting accuracy over the Northern Hemisphere.

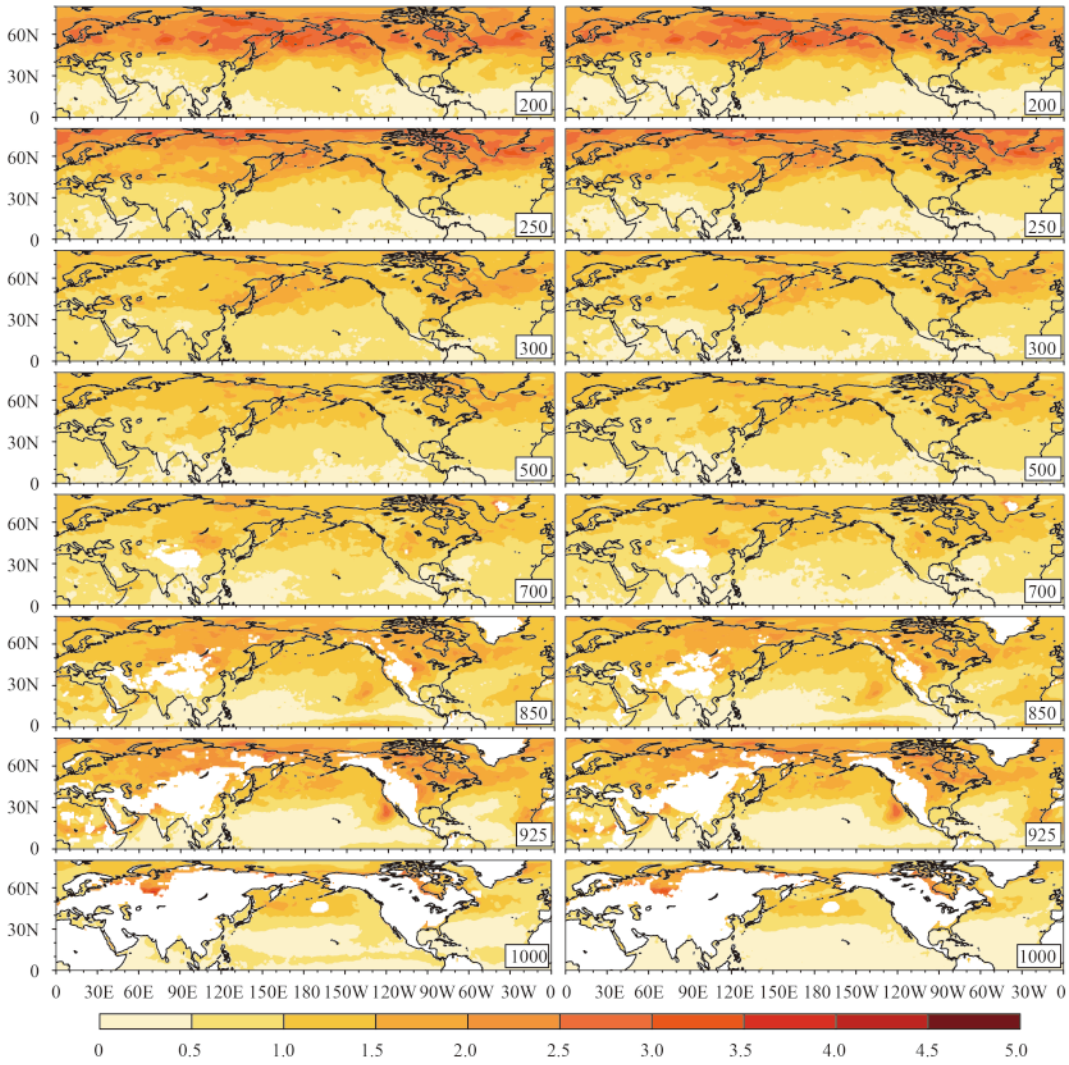

Fig. 9. As in Fig. 8, but for the 96-h forecast.

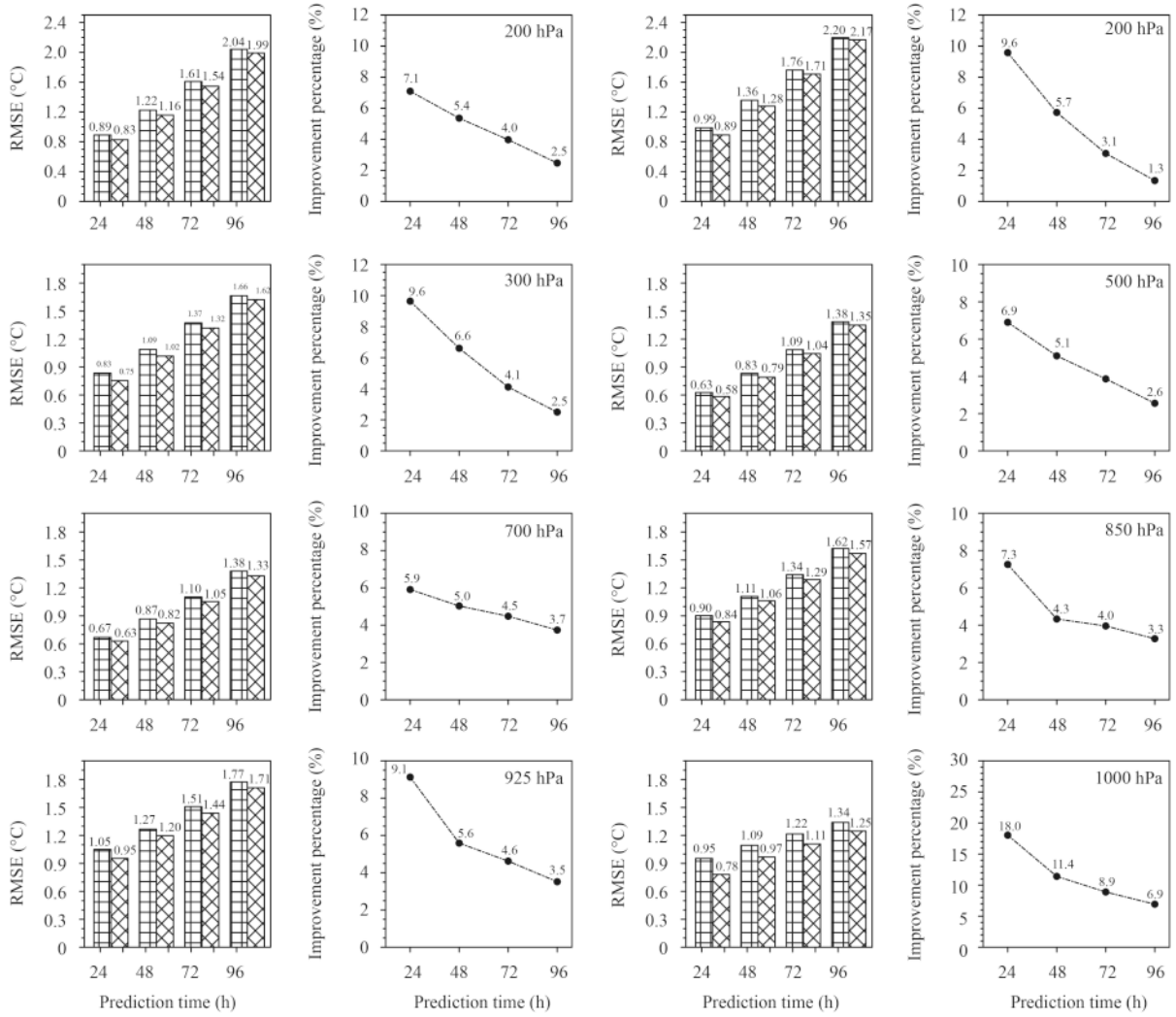

Fig. 10. As in Fig. 6, but for the pressure levels at 1000, 925, 850, 700, 500, 300, 250, and 200 hPa, over the Northern Hemisphere.

4. Conclusions and discussion

This study utilizes the inverse of expected error variance to determine the weight of individual ensemble members. To validate the method, the weights for ensemble members are determined based on TIGGE forecast data in the summer (1 June to 31 August) of 2012 over the Northern Hemisphere,with the NCEP-FNL data taken as the truth. Based on the weight of each ensemble member, the ensemble mean (i.e., the VWEM) is calculated for the summer (1 June to 31 August) of 2013, and the result is compared to that from the SEWEM method. The results show that the VWEM method has a lower AE and RMSE and has improved the prediction accuracy of the ensemble mean significantly. More specific conclusions are as follows.

(1) The VWEM method improves the forecasting accuracy considerably. The improvement percentages of 2-m temperature RMSEs are 15.5%, 14.3%, 13.3%, and 12.1% at the prediction lead-times of 24, 48, 72, and 96 h, respectively, over the Northern Hemisphere, among which the improvement percentage (> 24%) over the Tibetan Plateau and its surrounding areas is particularly notable.

(2) Compared to the SEWEM method, the VWEM method reduces the RMSE for 2-m temperature by almost 0.4°C at all prediction lead-times over the Northern Hemisphere. There is considerable improvement over the Tibetan Plateau and its surrounding areas. The absolute improvement is about 1.0°C.

(3) Smaller RMSEs of the ensemble mean are calculated by the VWEM method at the different pressure levels, as compared with those calculated by the SEWEM method. The VWEM method reduces the RMSE by almost the same at the same pressure level with increasing forecasting lead-time, though the im-

provement percentages are different.

Although in this study the VWEM method is applied only to temperature at prediction lead-times of 24, 48, 72, and 96 h over the Northern Hemisphere, the method can be, in principle, employed for any meteorological variable and at longer lead-times. It should also be pointed out that, for convenience, only the spatial variation weights are used in this study. That is, the weights do not update with time. In other words, the weight is the same at all times during the summer of 2013 in the VWEM method. Another weakness is that the NCEP-FNL data are taken as the truth to verify the model simulations. This may have an influence on the weights. For example, the NCEP forecasts have a large weight, as compared to the other members. As a result, the ensemble mean may be impacted by taking the NCEP-FNL data as the truth. In view of this, the next step in this line of research is to take into account a rolling updated weight, and use surface and sounding observations as the truth. In addition, more meteorological variables (e.g., geopotential height, relative humidity, wind speed and direction) will be employed and validated in future work. Besides, the VWEM method will be applied to longer prediction lead-times, such as 10 days or more. Moreover, pre-processing of bias correction for each member will be performed (Tao et al., 2014) and tests will be launched without some ensemble members with low weights, which may improve the ensemble mean (Zhi et al., 2012). It is hoped that this approach can be applied in operational ensemble prediction systems, and provide guidance to weather and climate prediction. It should be noted that much work has already been accomplished on multi-model weighting and analysis (e.g., Raftery et al., 2005; Weigel et al., 2008; Du and Zhou, 2011; Baran and Lerch, 2015), and thus a study in which these methods are compared will also be conducted in the near future.

Acknowledgements. The TIGGE datasets were provided by the ECMWF, and the NCEP-FNL data were from GDAS (the Global Data Assimilation System). We thank the three anonymous reviewers for providing constructive comments and suggestions, which greatly improved the quality of the paper.

REFERENCES

Baran, S., and S. Lerch, 2015: Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quart. J. Roy. Meteor. Soc.*, **141**, 2289–2299, doi: 10.1002/qj.2521.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, doi: 10.1038/nature14956.

Bougeault, P., Z. Toth, C. Bishop, et al., 2010: The THORPEX interactive grand global ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, doi: 10.1175/2010BAMS2853.1.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, doi: 10.1175/2008MWR2682.1.

Charron, M., G. Pellerin, L. Spacek, et al., 2009: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901, doi: 10.1175/2009MWR3187.1.

Chen, G. S., Z. Liu, and J. E. Kutzbach, 2014: Reexamining the barrier effect of the Tibetan Plateau on the South Asian summer monsoon. *Clim. Past*, **10**, 1269–1275, doi: 10.5194/cp-10-1269-2014.

Chen, J. Q., and S. Bordoni, 2014: Orographic effects of the Tibetan Plateau on the East Asian summer monsoon: An energetic perspective. *J. Climate*, **27**, 3052–3072, doi: 10.1175/JCLI-D-13-00479.1.

Cui, H. H., and X. F. Zhi., 2013: Multi-model ensemble forecasts of surface air temperature in the extended range using the TIGGE dataset. *Trans. Atmos. Sci.*, **36**, 165–173. (in Chinese)

Cui, M. C., M. Feng, S. M. Lian, et al., 2000: Evaluation of daily precipitation in China from ECMWF and NCEP reanalyses. *Chin. J. Ocean Limnol.*, **18**, 35–41, doi: 10.1007/BF02842539.

Descamps, L., C. Labadie, A. Joly, et al., 2014: RP PEA, the Météo-France short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1671–1685, doi: 10.1002/qj.2469.

Du, J., 2007: Uncertainty and Ensemble Forecasting. NOAA/NWS Science and Technology Infusion Lecture Series, 42. [Available online at http://www.nws.noaa.gov/ost/climate/STIP/uncertainty.htm].

Du, J., and B. B. Zhou, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Mon. Wea. Rev.*, **139**, 3284–3303, doi: 10.1175/MWR-D-10-05007.1.

Duan, A. M., G. X. Wu, Y. M. Liu, et al., 2012: Weather and climate effects of the Tibetan Plateau. *Adv. Atmos. Sci.*, **29**, 978–992, doi: 10.1007/s00376-012-1220-y.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi: 10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Froude, L. S. R., 2011: TIGGE: Comparison of the prediction of Southern Hemisphere extratropical cyclones by different ensemble prediction systems. *Wea. Forecasting*, **26**, 388–398, doi: 10.1175/2010WAF2222326.1.

He, C. F., X. F. Zhi, Q. L. You, et al., 2015: Multi-model ensemble forecasts of tropical cyclones in 2010 and 2011 based on the Kalman Filter method. *Meteor. Atmos. Phys.*, **127**, 467–479, doi: 10.1007/s00703-015-0377-1.

Huffman, G. J., R. F. Adler, B. Rudolf, et al., 1995: Global precipitation estimates based on a technique for combining satellite-based estimates, rain gauge analysis, and NWP model precipitation information. *J. Climate*, **8**, 1284–1295, doi: 10.1175/1520-0442(1995)008<1284:GPEBOA>2.0.CO;2.

Huffman, G. J., R. F. Adler, P. Arkin, et al., 1997: The Global Precipitation Climatology Project (GPCP) combined precipitation dataset. *Bull. Amer. Meteor. Soc.*, **78**, 5–20, doi: 10.1175/1520-0477(1997)078<0005:TGPCPG>2.0.CO;2.

Jewson, S., 2013: A simple method for eliminating double counting in multi-model ensemble forecasts. Working Paper. [Available online at http://cedadocs.badc.rl.ac.uk/961/].

Krishnamurti, T. N., A. D. Sagadevan, A. Chakraborty, et al., 2009: Improving multimodel weather forecast of monsoon rain over China using FSU superensemble. *Adv. Atmos. Sci.*, **26**, 813–839, doi: 10.1007/s00376-009-8162-z.

Krishnamurti, T. N., C. M. Kishtawal, D. W. Shin, et al., 2000a: Improving tropical precipitation forecasts from a multianalysis superensemble. *J. Climate*, **13**, 4217–4227, doi: 10.1175/1520-0442(2000)013<4217:ITPFFA>2.0.CO;2.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, et al., 2000b: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, doi: 10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, doi: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Leutbecher, M., and S. T. K. Lang, 2014: On the reliability of ensemble variance in subspaces defined by singular vectors. *Quart. J. Roy. Meteor. Soc.*, **140**, 1453–1466, doi: 10.1002/qj.2229.

Lin, C. Z., X. F. Zhi, Y. Han, et al., 2009: Multi-model super-ensemble forecasts of the surface temperature using the TIGGE data. *J. Appl. Meteor. Sci.*, **20**, 706–712. (in Chinese)

Liu, J. G., and Z. H. Xie, 2014: BMA probabilistic quantitative precipitation forecasting over the Huaihe basin using TIGGE multimodel ensemble forecasts. *Mon. Wea. Rev.*, **142**, 1542–1555, doi: 10.1175/MWR-D-13-00031.1.

Liu, X. D., and B. W. Dong, 2013: Influence of the Tibetan Plateau uplift on the Asian monsoon-arid environment evolution. *Chin. Sci. Bull.*, **58**, 4277–4291, doi: 10.1007/s11434-013-5987-8.

Magnusson, L., J.-R. Bidlot, S. T. K. Lang, et al., 2014: Evaluation of medium-range forecasts for hurricane sandy. *Mon. Wea. Rev.*, **142**, 1962–1981, doi: 10.1175/MWR-D-13-00228.1.

Molteni, F., R. Buizza, T. N. Palmer, et al., 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, doi: 10.1002/qj.49712252905.

Qi, L. B., H. Yu, P. Y. Chen, 2014: Selective ensemble-mean technique for tropical cyclone track forecast by using ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **140**, 805–813, doi: 10.1002/qj.2196.

Raftery, A. E., T. Gneiting, F. Balabdaoui, et al., 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi: 10.1175/MWR2906.1.

Smith, L. A., H. L. Du, E. B. Suckling, et al., 2014: Probabilistic skill in ensemble seasonal forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 1085–1100, doi: 10.1002/qj.2403.

Tao, Y. M., Q. Y. Duan, A. Z. Ye, et al., 2014: An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai River basin. *J. Hydrol.*, **519**, 2890–2905, doi: 10.1016/j.jhydrol.2014.04.040.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci.*, **365**, 2053–2075.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi: 10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the national meteorological center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, doi: 10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

van Schaeybroeck, B., and S. Vannitsem, 2014: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, doi: 10.1002/qj.2397.

Wang, H. X., and X. F. Zhi, 2015: Statistical downscaling research of precipitation forecast based on TIGGE multimodel ensemble. *J. Meteor. Sci.*, **35**, 430–437. (in Chinese)

Wang, A. H., and X. B. Zeng, 2012: Evaluation of multireanalysis products with in situ observations over the Tibetan Plateau. *J. Geophys. Res.*, **117**, D05102, doi: 10.1029/2011JD016553.

Wei, M. Z., C. Rowley, P. Martin, et al., 2014: The US Navy's RELO ensemble prediction system and its performance in the Gulf of Mexico. *Quart. J. Roy. Meteor. Soc.*, **140**, 1129–1149, doi: 10.1002/qj.2199.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134**, 241–260, doi: 10.1002/qj.210.

Xie, P. P., and A. Arkin, 1996: Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Climate*, **9**, 840–858, doi: 10.1175/1520-0442(1996)009<0840:AOGMPU>2.0.CO;2.

Ye, J., Y. He, F. Pappenberger, et al., 2014: Evaluation of ECMWF medium-range ensemble forecasts of precipitation for river basins. *Quart. J. Roy. Meteor. Soc.*, **140**, 1615–1628, doi: 10.1002/qj.2243.

Zhang, H. B., X. F. Zhi, J. Chen, et al., 2015: Study of the modification of multi-model ensemble schemes for tropical cyclone forecasts. *J. Trop. Meteor.*, **21**, 389–399. (in Chinese)

Zhao, L. N., H. Wu, F. Y. Tian, et al., 2010: Assessment of probabilistic precipitation forecasts for the Huaihe Basin using TIGGE data. *Meteor. Mon.*, **36**, 133–142.

Zhi, X. F., H. X. Qi, Y. Q. Bai, et al., 2012: A comparison of three kinds of multimodel ensemble forecast techniques based on the TIGGE data. *Acta Meteor. Sinica*, **26**, 41–51, doi: 10.1007/s13351-012-0104-5.

Zhi, X. F., X. D. Ji, J. Zhang, et al., 2013: Multimodel ensemble forecasts of surface air temperature and precipitation using TIGGE datasets. *Trans. Atmos. Sci.*, **36**, 257–266. (in Chinese)

Zhou, B. B., and J. Du, 2010: Fog prediction from a multimodel mesoscale ensemble prediction system. *Wea. Forecasting*, **25**, 303–322, doi: 10.1175/2009WAF2222289.1.

Zhu, J. S., F. Y. Kong, and H. C. Lei, 2012: A regional ensemble forecast system for stratiform precipitation events in northern China. Part I: A case study. *Adv. Atmos. Sci.*, **29**, 201–216, doi: 10.1007/s00376-011-0137-1.

Zhu, J. S., F. Y. Kong, and H. C. Lei, 2013: A regional ensemble forecast system for stratiform precipitation events in the northern China region. Part II: Seasonal evaluation for summer 2010. *Adv. Atmos. Sci.*, **30**, 15–28, doi: 10.1007/s00376-012-1043-x.

Tech & Copy Editor: Shuri LI
Language Editor: Colin SMITH