ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 33, NOVEMBER 2016, 1297–1305

Ensemble Mean Forecast Skill and Applications with the T213 Ensemble Prediction System

Sijia LI, Yuan WANG, Huiling YUAN*, Jinjie SONG, and Xin XU

Key Laboratory of Mesoscale Severe Weather, Ministry of Education, School of Atmospheric Sciences, Nanjing University, Nanjing 210023*, China*

(Received 19 June 2016; revised 27 July 2016; accepted 8 August 2016)

ABSTRACT

Ensemble forecasting has become the prevailing method in current operational weather forecasting. Although ensemble mean forecast skill has been studied for many ensemble prediction systems (EPSs) and different cases, theoretical analysis regarding ensemble mean forecast skill has rarely been investigated, especially quantitative analysis without any assumptions of ensemble members. This paper investigates fundamental questions about the ensemble mean, such as the advantage of the ensemble mean over individual members, the potential skill of the ensemble mean, and the skill gain of the ensemble mean with increasing ensemble size. The average error coefficient between each pair of ensemble members is the most important factor in ensemble mean forecast skill, which determines the mean-square error of ensemble mean forecasts and the skill gain with increasing ensemble size. More members are useful if the errors of the members have lower correlations with each other, and vice versa. The theoretical investigation in this study is verified by application with the T213 EPS. A typical EPS has an average error coefficient of between 0.5 and 0.8; the 15-member T213 EPS used here reaches a saturation degree of 95% (i.e., maximum 5% skill gain by adding new members with similar skill to the existing members) for 1–10-day lead time predictions, as far as the mean-square error is concerned.

Key words: ensemble mean, forecast skill, ensemble size, ensemble prediction system, saturation degree

Citation: Li, S. J., Y. Wang, H. L. Yuan, J. J. Song, and X. Xu, 2016: Ensemble mean forecast skill and applications with the T213 ensemble prediction system. *Adv. Atmos. Sci.*, **33**(11), 1297–1305, doi: 10.1007/s00376-016-6155-2.

1. Introduction

The principle of combining forecasting outputs from different models and members into an ensemble was proposed several decades ago (Sanders, 1963; Epstein, 1969; Leith, 1974) and has been widely employed in meteorology and other fields since the 1990s, especially the arithmetic average of all ensemble members, i.e., the ensemble mean. From an experimental perspective, it is well known that the ensemble mean often outperforms its individual members in operational forecasts (Vislocky and Fritsch, 1995; Fritsch et al., 2000). Recently, other complex methods have been developed to construct unequally weighted or bias-corrected ensembles instead of the arithmetic mean, such as linear regressions (Krishnamurti et al., 1999, 2000), nonlinear regressions (Hamill et al., 2008), Bayesian averages (Raftery et al., 2005; Vrugt et al., 2006), artificial neural networks (Yuan et al., 2007), and time-varying weighted bias correction methods (Hashino et al., 2007; Bohn et al., 2010). The improvements in the ensemble mean due to the application of these statistical methods vary on a case-by-case basis and are not stable due to the lack of a sufficient number of samples (Weisheimer et al., 2009). In fact, persistence in the relative skills of members is required to use complex weighting combination methods, except for simple arithmetic averaging (Reifen and Toumi, 2009). Therefore, the arithmetic ensemble mean remains one of the most effective methods in operational forecasts for many cases (Najafi and Moradkhani, 2016).

From a theoretical perspective, the pioneering work of Leith (1974) first examined the potential skill of Monte Carlo forecasts and found that the sample mean could better estimate the real state in comparison with conventional single forecasts. This indicated that the improvement of such a Monte Carlo ensemble in terms of the mean-square skill was a consequence of the optimal filtering nature of the procedure. Additionally, several recent studies have attempted to reveal the essence of the forecast skill of an ensemble mean. Hagedorn et al. (2005) argued that the success of multimodel ensemble means was mainly due to error cancellation and nonlinearity of skill score metrics. Weigel et al. (2008) further found that a "poorer" member can also contribute to the skill of an ensemble mean. There are also other studies that have examined the advantages of using an ensemble mean by studying the relationships between ensemble members. For

* Corresponding author: Huiling YUAN
Email: yuanhl@nju.edu.cn

example, the members that have higher skills and are less dependent on each other have been suggested for an ensemble prediction system (EPS) to achieve the best ensemble mean skill (Yoo and Kang, 2005). Jeong and Kim (2009) demonstrated that neither the equally nor unequally weighted mean method could effectively improve the forecast skill if significant correlations exist between the members. However, their study only targeted two-member combinations and assumed that the two members were unbiased. Winter and Nychka (2010) conceptually indicated that the ensemble mean can outperform the best individual member if the forecasting outputs of the ensemble members are markedly different from each other. However, the relationship between the ensemble mean skill and the correlation of the ensemble members has rarely been quantitatively deduced without any assumptions.

Another important issue is the role of the ensemble size on the performance of EPSs. In previous studies, it has been concluded that a limited number of ensemble members is sufficient to achieve a saturated skill (Houtekamer and Derome, 1995; Deque, 1997; Buizza and Palmer, 1998). Du et al. (1997) indicated that an ensemble size of 8–10 can account for a near 90% reduction in the RMSEs of ensemble mean precipitation forecasts. Clark et al. (2011) revealed that the skill gain decreases with increasing ensemble size. Ma et al. (2012) found that more members are required to increase the forecast skill, especially for long-range forecasts, although the improvements were found to be insignificant beyond 20 members when measured by deterministic metrics. All of the above research was based on experimental studies. From a theoretical perspective, Richardson (2001) discussed the impact of the ensemble size on probabilistic forecasts in terms of Brier scores, reliability diagrams and potential economic value, and found that for different metrics, the sufficient ensemble size is different. However, the impact of the ensemble size on the mean-square error (MSE), which is one of the most commonly used deterministic metrics, has rarely been discussed in a theoretic context.

This study aims to investigate the potential forecast skill of the ensemble mean, including the optimum ensemble mean and its superiority over its individual members, and the impact of ensemble size, without specific assumptions regarding the ensemble members. The theoretical analyses related to the fundamental questions of the ensemble mean are described in section 2. Experimental studies based on the China Meteorological Administration (CMA) T213L31 EPS (hereafter, T213 EPS) are presented in section 3. Section 4 gives a summary and discussion.

2. Theoretical analysis of the ensemble mean

2.1. *Ensemble mean skill*

For a finite number of data points K, the forecasts from M ensemble members $(\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_M)$ can be combined to construct an ensemble mean $\boldsymbol{F}^*$, where $\boldsymbol{F}_i = (F_{i,1}, F_{i,2}, \ldots, F_{i,K})$, $i = 1, 2, \ldots, M$, and $\boldsymbol{F}^* = (F_1^*, F_2^*, \ldots, F_K^*)$. $F_{i,k}$ denotes the forecast at the kth data point predicted by the ith member, and $\boldsymbol{T} = (T_1, T_2, \ldots, T_K)$ denotes the corresponding validation values.

Let $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_M$ and $\boldsymbol{E}^*$ denote the errors for each member and the ensemble mean, respectively, where $\boldsymbol{E}_i = (e_{i,1}, \ldots, e_{i,K})$ and $e_{i,k} = F_{i,k} - T_k$. The errors $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_M$, $\boldsymbol{E}^*$ can be viewed as random variables with expectations $\overline{\boldsymbol{E}_1}, \overline{\boldsymbol{E}_2}, \ldots, \overline{\boldsymbol{E}_M}, \overline{\boldsymbol{E}^*}$. Thus, $\overline{\boldsymbol{E}_i} = (1/K)\sum_{k=1}^{K} e_{i,k}$.

The forecast error is expressed by the MSE:

$$R_i^2 = \frac{1}{K}\left[\sum_{k=1}^{K}(F_{i,k} - T_k)^2\right] = \frac{1}{K}\sum_{k=1}^{K} e_{i,k}^2, \quad i = 1, 2, \ldots, M \,. \tag{1}$$

For the ensemble mean $\boldsymbol{E}^* = (e_1^*, e_2^*, \ldots, e_K^*)$, there exists

$$\begin{aligned} e_k^* &= \frac{1}{M}\sum_{i=1}^{M} e_{i,k}, \quad k = 1, 2, \ldots, K \,, \\ R^{*2} &= \frac{1}{K}\sum_{k=1}^{K} e_k^{*2} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{M}\sum_{i=1}^{M} e_{i,k}\right)^2 \\ &= \frac{1}{M^2}\left[\sum_{i=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K} e_{i,k}^2\right) + 2\sum_{\substack{i,j=1 \\ i\neq j}}^{M}\left(\frac{1}{K}\sum_{k=1}^{K} e_{i,k}e_{j,k}\right)\right] . \end{aligned} \tag{2}$$

Let the following denote the error covariance between the ith and jth members:

$$R_{i,j} = \frac{1}{K}\sum_{k=1}^{K} e_{i,k}e_{j,k} \,. \tag{3}$$

The MSE of the ensemble mean can be calculated as

$$R^{*2} = \frac{1}{M^2}\left[\sum_{i=1}^{M} R_i^2 + 2\sum_{\substack{i,j=1 \\ i\neq j}}^{M} R_{i,j}\right] . \tag{4}$$

The errors of all ensemble members can be represented by a matrix $\boldsymbol{E}$:

$$\boldsymbol{E} = \begin{bmatrix} R_1^2 & \cdots & R_{1,M} \\ \vdots & \ddots & \vdots \\ R_{M,1} & \cdots & R_M^2 \end{bmatrix}_{M\times M} . \tag{5}$$

The MSE of the ensemble mean in Eq. (4) is equal to the average of the elements of matrix $\boldsymbol{E}$ in Eq. (5). The matrix $\boldsymbol{E}$ is symmetrical because $R_{i,j} = R_{j,i}$. The diagonal elements indeed represent the forecast skill of the individual ensemble members according to Eq. (1), whereas the remaining elements $R_{i,j}$ represent the relationship between the errors of any two members. This actually reveals the mathematical essence of the ensemble mean in that the skill of the ensemble mean depends on both the skills of the individual ensemble members and the relationship between the errors of any two members. This result is the generalization of Jeong and Kim (2009) because no assumptions were made in Eq. (4).

If we want to add a new member $\boldsymbol{F}_{M+1}$ to the already existing M-member ensemble, the MSE of the new ensemble is

equal to the average elements of matrix $\boldsymbol{E}_{M+1}$:

$$\boldsymbol{E}_{M+1} = \begin{bmatrix} R_1^2 & \cdots & R_{1,M} & R_{1,M+1} \\ \vdots & \ddots & \vdots & \vdots \\ R_{M,1} & \cdots & R_M^2 & R_{M,M+1} \\ R_{M+1,1} & \cdots & R_{M+1,M} & R_{M+1}^2 \end{bmatrix}_{(M+1)\times(M+1)} = \begin{bmatrix} \boldsymbol{E}_M & & R_{1,M+1} \\ & & \vdots \\ R_{M+1,1} & \cdots & R_{M+1}^2 \end{bmatrix}. \quad (6)$$

The $(M+1)$-member ensemble can outperform the already existing M-member ensemble if and only if the average elements of $\boldsymbol{E}_{M+1}$ are smaller than $\boldsymbol{E}_M$. This means that the average of the newly added elements in Eq. (6) should be smaller than the average elements of $\boldsymbol{E}_M$:

$$\frac{1}{2M+1}\left(R_{M+1}^2 + 2\sum_{i=1}^{M} R_{i,M+1}\right) < R^{*2} . \quad (7)$$

Equation (7) gives the essential and sufficient conditions in which a new member can enhance the skill of the already existing ensemble mean. Instead of simply having better skill, the newly added members should be less correlated with the already existing ensemble members because of the weights in Eq. (7). This can explain why a "poorer" member can still enhance the skill of the ensemble mean, which was discovered by Weigel et al. (2008). Otherwise, even if the new member has a higher skill than the existing members, it can still decrease the ensemble mean if it is highly correlated with the already existing ensemble members.

For different i and j, we have the following:

$$e_{i,k}e_{j,k} \leqslant \frac{1}{2}(e_{i,k}^2 + e_{j,k}^2) .$$

Moreover, the following is also valid:

$$R_{i,j} = \frac{1}{K}\sum_{k=1}^{K} e_{i,k}e_{j,k} \leqslant \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}(e_{i,k}^2 + e_{j,k}^2) = \frac{1}{2}(R_i^2 + R_j^2) . \quad (8)$$

From Eqs. (4) and (8), the following can be obtained:

$$\begin{aligned} R^{*2} &\leqslant \frac{1}{M^2}\left[\sum_{i=1}^{M} R_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{M}(R_i^2 + R_j^2)\right] \\ &= \frac{1}{M^2}\left[\sum_{i=1}^{M} R_i^2 + (M-1)\sum_{i=1}^{M} R_i^2\right] \\ &= \frac{1}{M}\sum_{i=1}^{M} R_i^2 . \end{aligned} \quad (9)$$

Equation (9) demonstrates that the forecast skill of the ensemble mean always outperforms the average skill of the individual members. Thus, the ensemble mean can be used to avoid choosing the "poorer" single members if the relative performance of the individual members or the best member is unknown. This explains why the ensemble mean can often achieve a satisfactory skill in practice.

Let $R_{\min}^2 = \min(R_1^2, \ldots, R_M^2)$ denote the MSE of the best ensemble member. Moreover, let the following denote the average of the individual MSE R_i^2:

$$U = \frac{1}{M}\sum_{i=1}^{M} R_i^2 . \quad (10)$$

The average of all possible $R_{i,j}$ $(i \neq j)$ can be expressed as

$$L = \frac{2}{M(M-1)}\sum_{\substack{i,j=1 \\ i\neq j}}^{M} R_{i,j} . \quad (11)$$

From Eqs. (4), (10) and (11), the MSE of the ensemble mean can be written as

$$R^{*2} = \frac{1}{M}U + \left(1 - \frac{1}{M}\right)L . \quad (12)$$

As a result, the ensemble mean outperforms the best individual member if and only if

$$R_{\min}^2 > \frac{1}{M}U + \left(1 - \frac{1}{M}\right)L . \quad (13)$$

Equation (12) can also be explained in terms of the matrix $\boldsymbol{E}$ in Eq. (5) because U represents the average of the diagonal elements of $\boldsymbol{E}$, whereas L represents the average of all other elements of $\boldsymbol{E}$.

Equation (13) gives the sufficient and essential conditions in which the ensemble mean can achieve a higher skill than the best individual member, which occurs only under specific conditions, i.e., the members have similar skills and lower error covariances. This result is consistent with previous studies (Yoo and Kang, 2005; Jeong and Kim, 2009; Winter and Nychka, 2010), and it further indicates that the ensemble mean cannot outperform the best individual member if the members are highly correlated (larger L) or there are distinctively poorer members (noticeable increase of U). It can also explain why the multimodel ensemble occasionally cannot outperform its best individual model in numerical weather predictions (Hagedorn et al., 2012) if the individual models operated by different centers are highly correlated or the best model performs distinctively better than the other models.

2.2. *Role of ensemble size*

From Eqs. (9), (10) and (12), the following is valid:

$$U \geqslant R^{*2} \geqslant L , \quad (14)$$

which shows that U and L can be treated as upper and lower bounds of the MSE of the ensemble mean.

The MSE of the ensemble mean is equal to a weighted mean of U and L [Eq. (12)]. When the ensemble size M increases, the weight of the larger term U in Eq. (12) decreases, whereas the weight of the smaller term L increases. As a result, the error correlation for each pair of ensemble members becomes the main factor that determines the forecast skill of

the ensemble mean. If U and L remain constant with increasing ensemble size and the newly added members have similar attributes to the already existing members, the MSE of the ensemble mean should decrease toward its lower bound L and reach a saturated skill level.

Specifically, Eq. (12) can be simplified to

$$\lim_{M\to\infty} R^{*2} = L \,. \tag{15}$$

When $M \to \infty$, the lower bound L exactly represents the potential skill of the ensemble mean with increasing ensemble size.

Let ρ conceptually express the average error correlation coefficients between each pair of members:

$$\rho = \frac{L}{U} \,. \tag{16}$$

The parameter ρ describes similarity among ensemble members. Larger ρ implies the members are similar to each other. It is clear that $\rho \leqslant 1$.

From Eqs. (12) and (16), the MSE of the ensemble mean compared with its individual members R^{*2}/U can be written as a function of ρ and M:

$$\frac{R^{*2}}{U} = \frac{1}{M} + \left(1 - \frac{1}{M}\right) . \tag{17}$$

Obviously, $\rho \leqslant R^{*2}/U \leqslant 1$. When the ensemble size M increases, R^{*2}/U saturates to its lower bound ρ. The effect of the ensemble mean compared with its individual members is dependent on the average error coefficients and the ensemble size (Fig. 1). If the ensemble size M is sufficient, ρ exactly determines the effect of the ensemble mean. Smaller ρ leads to a better ensemble mean compared with individual members, which implies that the errors of the members should have lower correlations with each other to improve an ensemble mean.

The "saturation degree" can be defined to describe the relative distance between the MSE of the ensemble mean and its potential skill L:

$$S = \left(1 - \frac{R^{*2} - L}{L}\right) \times 100\% \,. \tag{18}$$

The saturation degree S increases when the ensemble size increases and the upper bound of S is 100%. By combining Eqs. (12) and (18), the saturation degree S can be simplified to

$$S = \left(1 - \frac{1-\rho}{M\rho}\right) \times 100\% \,. \tag{19}$$

Equation (19) can be rewritten as

$$M_{\text{saturate}} = \frac{1-\rho}{(1-S)\rho} \,. \tag{20}$$

Equation (20) implies that the minimal ensemble size to reach a given saturated skill is determined by the error correlation coefficients between each pair of ensemble members. Fewer members are required for a larger ρ, and vice versa (Table 1). When $\rho \to 0$, which implies that $L \to 0$ and the members are independent of each other, the skill of the ensemble mean can be effectively improved with increasing ensemble size. Conversely, when $\rho \to 1$, which implies that the individual

Table 1. The minimum ensemble sizes required to reach saturation degrees of 80%, 90%, 95% and 99%, according to Eq. (20).

	Saturation degree S			
ρ	80%	90%	95%	99%
0.1	45	90	180	900
0.2	20	40	80	400
0.3	12	24	47	234
0.4	8	15	30	150
0.5	5	10	20	100
0.6	4	7	14	67
0.7	3	5	9	43
0.8	2	3	5	25
0.9	1	2	3	12

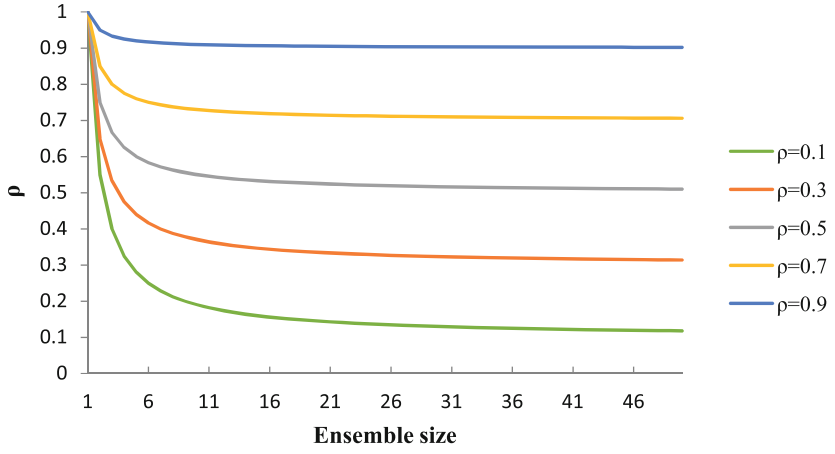

Fig. 1. MSE of the ensemble mean compared to its individual members R^{*2}/U, as a function of ρ and M, according to Eq. (17).

members are highly dependent, increasing the ensemble size is ineffective, and the improvement in the ensemble mean is negligible compared with the single members.

3. Application with the T213 EPS

The T213 EPS (Su et al., 2014) forecasts, which are provided by the CMA, have been archived in the TIGGE (Bougeault et al., 2010) database. The breeding initial perturbation method has been applied to the T213L31 (~60 km and 31 vertical levels) global spectral model (Chen et al., 2004; Wang et al., 2008) to generate 15 ensemble members, including one control run and seven pairs of perturbed members. This study uses the daily forecasts and corresponding analysis (validation) data from the T213 EPS over the Northern Hemisphere in 2008; the data have a $1^\circ \times 1^\circ$ output grid and 1–10-day lead time.

The MSE of the 500 hPa geopotential height, 850 hPa temperature and specific humidity, and 200 hPa wind speed (Fig. 2) shows that for a 1–3-day lead time, the ensemble mean of the 15 members performs slightly better than the control run and the average MSE of all its members U. With increasing lead time, the advantage of the ensemble mean becomes increasingly more significant for medium-range forecasts (4–10 days), despite the analysis field favoring the control run. Although the average MSE of the individual members U is appreciably larger, the ensemble mean outperforms the control run and is close to its lower bound L because the smaller L has a weight exceeding 90% (14/15) to determine R^{*2} according to Eq. (12) for the 15-member T213 EPS. With increasing lead time, the MSE of the individual members (including the control run) increases rapidly, whereas the increase in L is relatively slow. As a result, the error correlation coefficients between the ensemble members decrease when the lead time increases [Fig. 3; Eq. (16)]. This can explain why the advantage of the ensemble mean is more significant in medium-range forecasts than in short-range predictions.

The relationship between the forecast skill of the ensemble mean and ensemble size is also explored. There are $15!/[(15-i)!i!]$ choices to select i members from the 15-member ensemble. Among these choices, the best choice for each i is selected based on the lowest MSE of the ensemble mean. For the short-range forecasts (1–3 days), the skill (Fig. 4) of the best ensemble mean is barely improved by increasing the ensemble size. For the medium-range forecasts (4–10 days), the MSE of the ensemble mean rapidly decreases when the ensemble size increases, and the skill of the ensemble mean gradually becomes marginal and saturated.

The terms in Eq. (12) vary with the ensemble size. For example, with a 10-day lead time (Fig. 5), both U and L remain constant with increasing ensemble size, whereas the MSE of the ensemble mean decreases due to the change in the weight in Eq. (12). When the ensemble size M increases, the weight for the smaller term L increases toward 1, whereas the weight for the larger term U decreases toward 0. This indicates that the skill of the ensemble mean increases with increasing ensemble size and the ensemble mean skill becomes saturated because the weight change becomes smaller with a large M.

In this case, for the 1–10-day forecasts of the different fields of interest, such as the 500 hPa geopotential height, the 850 hPa temperature and specific humidity, and the 200 hPa wind speed, the parameter ρ varies between 0.5 and 0.8 (Fig. 3). The ensemble size required for a saturated ensemble mean

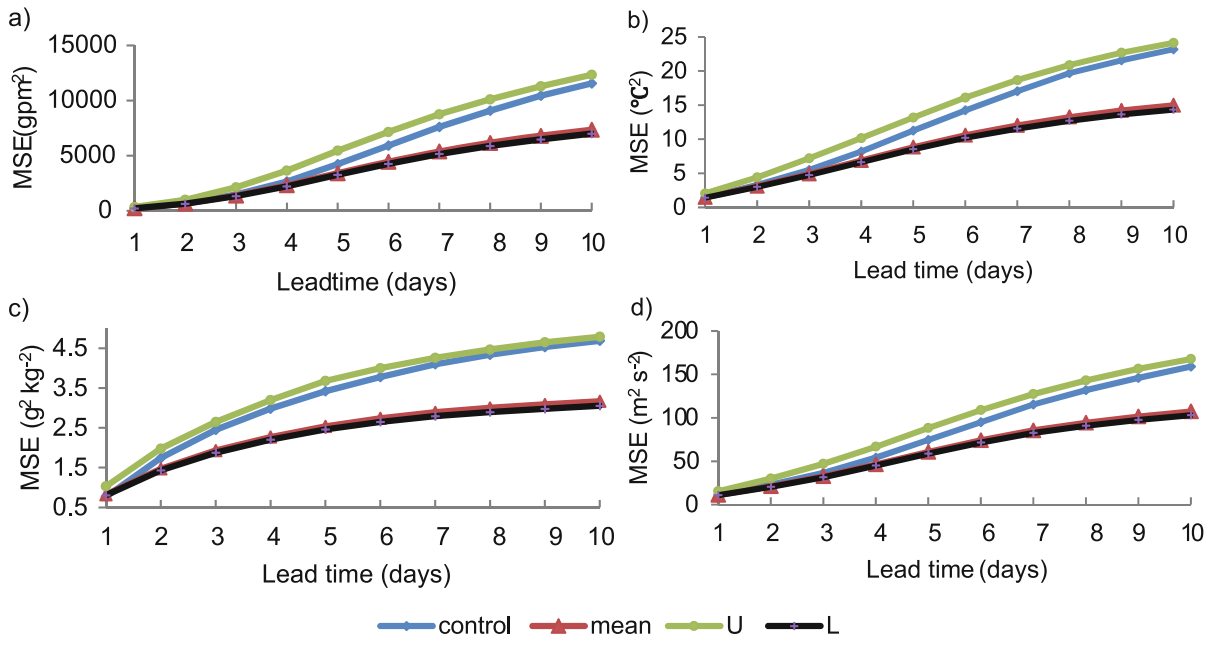

Fig. 2. MSE of the control run, the ensemble mean of the T213 EPS, and the upper and lower bounds U and L for 1–10-day lead times: (a) 500 hPa geopotential height; (b) 850 hPa temperature; (c) 850 hPa specific humidity; (d) 200 hPa wind speed.

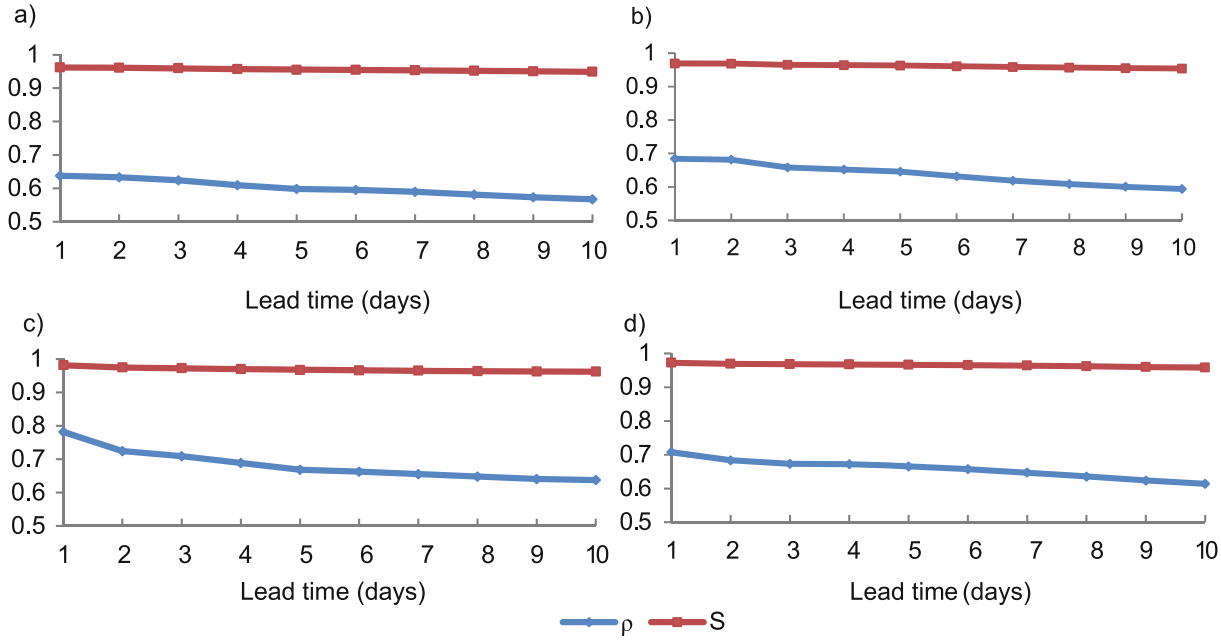

Fig. 3. The average error correlation coefficients for each pair of ensemble members ρ and the saturation degree S of the T213 EPS for 1–10-day lead times: (a) 500 hPa geopotential height; (b) 850 hPa temperature; (c) 850 hPa specific humidity; (d) 200 hPa wind speed.

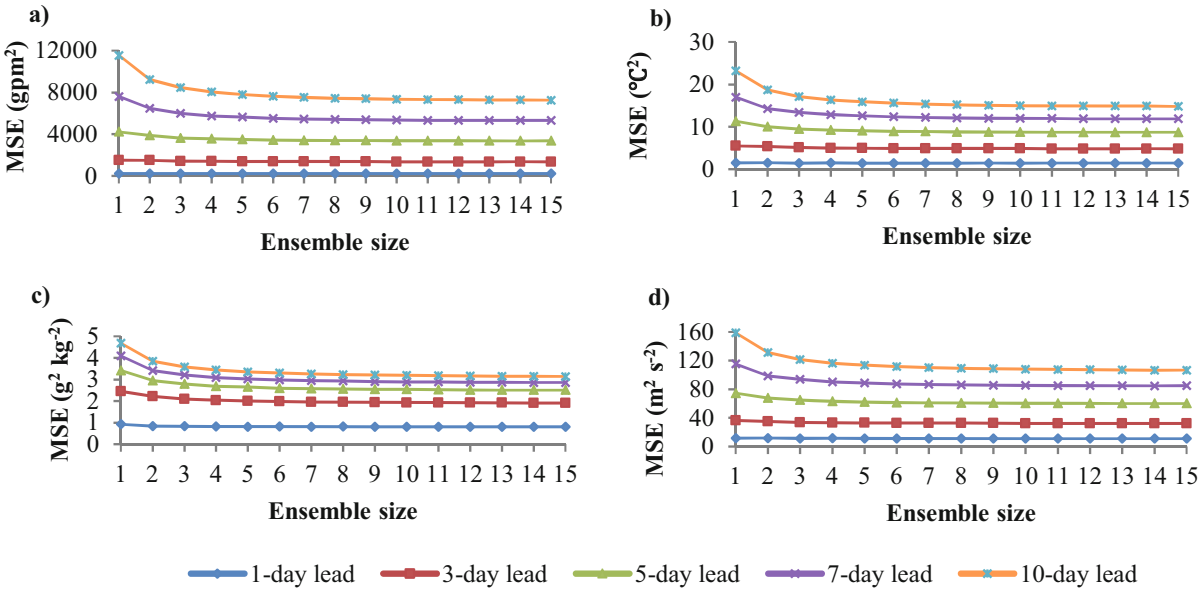

Fig. 4. MSE of the ensemble mean, as a function of the ensemble size, for lead times of 1, 3, 5, 7 and 10 days: (a) 500 hPa geopotential height; (b) 850 hPa temperature; (c) 850 hPa specific humidity; (d) 200 hPa wind speed.

can be deduced from Eq. (20) and Table 1. For different meteorological elements, the saturated ensemble size is different (Fig. 6). More members are required for medium-range forecasts than short-range predictions. For the four meteorological elements considered in this paper, 2–4 members can achieve a saturation degree of 80%, 3–8 members can reduce the MSE of the ensemble mean by 90%, and 6–16 members are enough to obtain a saturation degree of 95%. This can explain the previous results of Du et al. (1997) and Ma et al. (2012).

The already existing 15-member T213 EPS can reach a saturation degree of 95% for 1–10-day lead times (Fig. 3), except for the 10-day lead time of the 500 hPa geopotential height predictions, as far as the MSE is concerned. For the short-range predictions, the saturation degree is even higher. This implies that if new members are added to the 15-member T213 EPS, the MSE of the ensemble mean can only be reduced by 5%, unless the skills and error covariances of in-

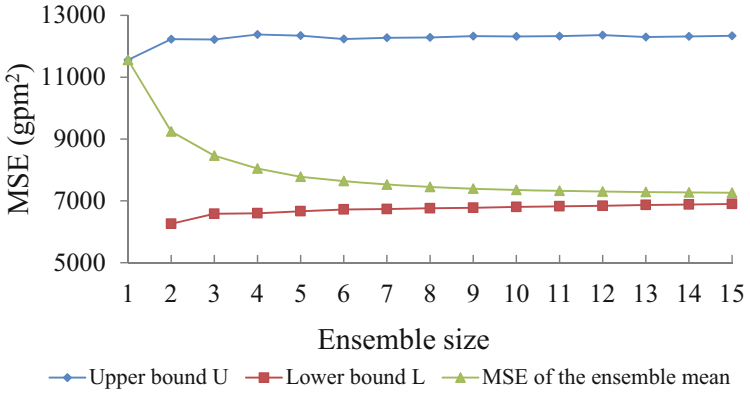

Fig. 5. MSE of the 500 hPa geopotential height for the ensemble mean R^{*2}, and its two factors U and L, as a function of the ensemble size M, for a lead time of 10 days (units: gpm^2).

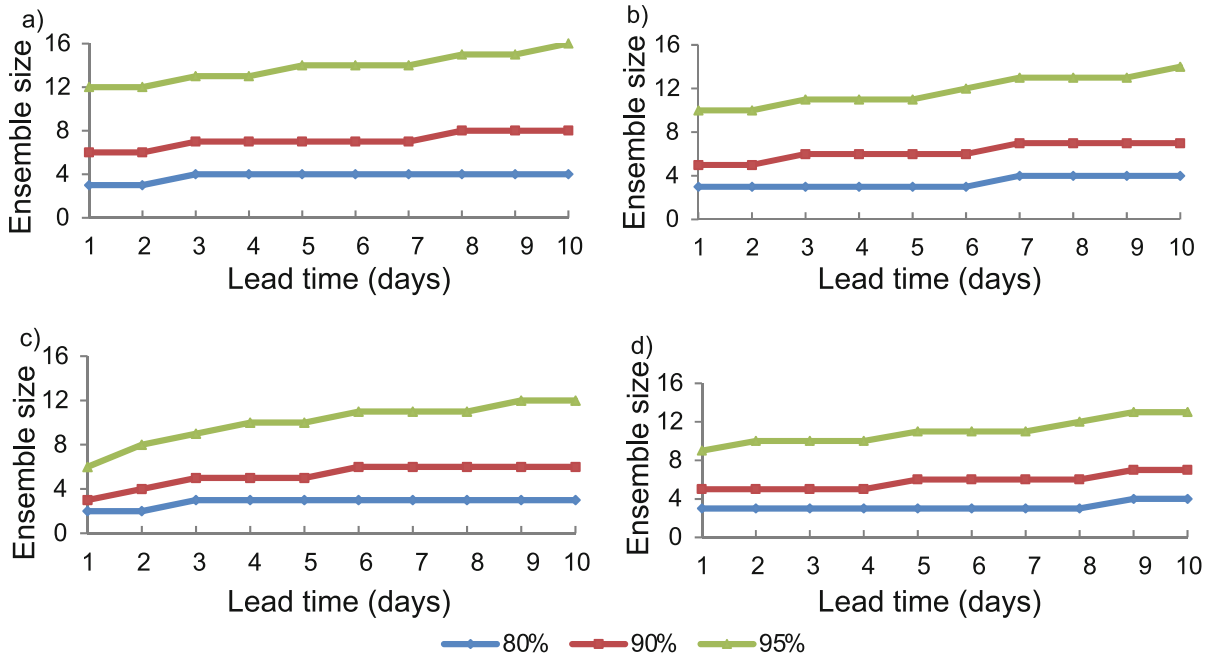

Fig. 6. The minimum ensemble sizes required to reach saturation degrees of 80%, 90% and 95%: (a) 500 hPa geopotential height; (b) 850 hPa temperature; (c) 850 hPa specific humidity; (d) 200 hPa wind speed.

dividual members are significantly improved. Note that the correlations between existing members of T213 EPS are high, which limits the skill gain of the ensemble mean when adding more ensemble members. For a better configured EPS that has lower correlations among the ensemble members, more ensemble mean skill is expected to be gained from more members.

4. Summary and discussion

Ensemble methods, especially the arithmetic mean, have been widely used in weather and climate forecasting. This paper set out to reveal the rationale behind the success of using ensemble means. The ensemble mean cannot always outperform the best single member, although it has a better skill than the average skill of all individual members. The skill of the ensemble mean not only depends on the skills of individual members, but even more so on the error covariances between each pair of ensemble members. This suggests that the best approach is to choose ensemble members that have lower error covariances with each other, to achieve a better ensemble mean skill.

It is inappropriate to blindly add new members into an already existing ensemble. A greater ensemble size does not necessarily yield higher skill. Even if a new member has a higher skill, it can still decrease the ensemble skill if it is highly correlated with the already existing ensemble mem-

bers. In addition, the ensemble mean skill tends to saturate toward its potential skill when the ensemble size increases under the condition that the newly added members have similar attributes to the already existing ensemble members. This also indicates that increasing ensemble size will benefit the ensemble mean more when the added members have lower covariances with existing members.

The average error coefficient between individual ensemble members is the most important factor to determine the ensemble skill. It not only determines the effect of the ensemble mean compared with individual members, but also the potential skill and the saturation degree of the ensemble mean. More members are useful if the errors of the members have lower correlations with each other, and vice versa.

The T213 EPS forecasts confirm the above theoretical results. The ensemble mean of the T213 EPS outperforms its control run, especially for medium-range forecasts, because the error covariances between each pair of ensemble members are lower than the MSEs of the individual members. The skill of the ensemble mean can be improved by increasing the ensemble size for medium-range forecasts, which saturates gradually, under the condition that the perturbed members have similar attributes to each other. However, the ensemble mean skill of the short-range forecasts saturates quickly with a small ensemble size.

For an ensemble that has an average error correlation coefficient that varies between 0.5 and 0.8, 15 members already reach a saturated ensemble mean. The 15-member T213 EPS can reach a saturation degree of 95% for 1–10-day lead time predictions, as far as the MSE is concerned. For short-range forecasts, the saturated ensemble size is even smaller. This can also be attributed to a greater correlation between ensemble members in short-range forecasts. The already existing ensemble can barely be improved by simply adding new members that have similar attributes to the already existing members. The T213 EPS members show high correlations, and for this reason its ensemble mean skill saturates quickly at around 10 members, especially for specific humidity forecasts at shorter lead times. Therefore, efforts should be made to reduce the correlations among the ensemble members to benefit from more members. In addition, we only examine the ensemble mean skill score in a deterministic sense in this study; we do not address the probabilistic forecasting aspect. It is very likely that probabilistic forecasting skill can further benefit from more ensemble members, even when the ensemble mean skill score ceases to improve by adding additional members. Further research from the probabilistic forecasting perspective is still needed.

In this paper, the MSE is used as the metric to evaluate the ensemble mean. For different metrics, the theoretical frameworks are different. Theoretical analyses based on other metrics and the internal relationship between different metrics still requires further study.

Although these theoretical analyses in this study focus on the ensemble mean with equal weights, they can also be generalized to an unequally weighted ensemble mean. Obviously, to obtain a better weighted mean, larger weights should be assigned to the members with higher skill. Further research is needed on weight setting methods.

This study is based on the EPS of a single center (the CMA); the error covariances between the outputs of different centers in the THORPEX TIGGE data may improve the skill of the ensemble mean. Research on multi-center models requires further study.

Acknowledgements. The T213 EPS data were obtained from the ECMWF TIGGE portal (http://tigge-portal.ecmwf.int/). This work was supported by the National Basic Research (973) Program of China (Grant No. 2013CB430106), the R&D Special Fund for Public Welfare Industry (Meteorology) (Grant Nos. GYHY201306002 and GYHY201206005), the National Natural Science Foundation of China (Grant Nos. 40830958 and 41175087), the Jiangsu Collaborative Innovation Center for Climate Change, and the High Performance Computing Center of Nanjing University. The authors would like to thank the two anonymous reviewers, the editor(s) and the editorial office for carefully reviewing the manuscript and providing us with constructive comments and valuable suggestions for revision.

REFERENCES

Bohn, T. J., M. Y. Sonessa, and D. P. Lettenmaier, 2010: Seasonal hydrologic forecasting: Do multimodel ensemble averages always yield improvements in forecast skill? *J. Hydrometeorol.*, **11**(6), 1358–1372.

Bougeault, P., and Coauthors, 2010: The THORPEX interactive grand global ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518.

Chen, Q. Y., M. M. Yao, and Y. Wang, 2004: A new generation of operational medium-range weather forecast model T213L31 in National Meteorological Center. *Meteorological Monthly*, **30**(10), 16–21. (in Chinese)

Clark, A. J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418.

Deque, M., 1997: Ensemble size for numerical seasonal forecasts. *Tellus A*, **49**, 74–86.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Fritsch, J. M., J. Hilliker, J. Ross, and R. L. Vislocky, 2000: Model consensus. *Wea. Forecasting*, **15**, 571–582.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A*, **57**(3), 219–233.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble re-

forecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences*, **11**(2), 939–950.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**, 2181–2196.

Jeong, D., and Y. O. Kim, 2009: Combining single-value streamflow forecasts—A review and guidelines for selecting techniques. *J. Hydrol.*, **377**(3–4), 284–299.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**(5433), 1548–1550.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**(23), 4196–4216.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Ma, J. H., Y. J. Zhu, R. Wobus, and P. X. Wang, 2012: An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.*, **29**, 782–794, doi: 10.1007/s00376-012-1249-y.

Najafi, M. R., and H. Moradkhani, 2016: Ensemble combination of seasonal streamflow forecasts. *Journal of Hydrologic Engineering*, **21**(1), 04015043.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**(5), 1155–1174.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36**, L13704, doi: 10.1029/2009GL038082.

Richardson, D. S., 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.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.

Su, X., H. L. Yuan, Y. J. Zhu, Y. Luo, and Y. Wang, 2014: Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. *J. Geophys. Res. Atmos.*, **119**, 7292–7310.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.*, **76**(7), 1157–1164.

Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Y. Duan, and B. A. Robinson, 2006: Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophys. Res. Lett.*, **33**(19), L19817, doi: 10.1029/2006GL027126.

Wang, Y., H. Qian, J.-J. Song, and M.-Y. Jiao, 2008: Verification of the T213 global spectral model of China National Meteorology Center over the East-Asia area. *J. Geophys. Res.*, **113**, D10110, doi: 10.1029/2007JD008750.

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**(630), 241–260.

Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21711, doi: 10.1029/2009GL040896.

Winter, C. L., and D. Nychka, 2010: Forecasting skill of model averages. *Stochastic Environmental Research and Risk Assessment*, **24**(5), 633–638.

Yoo, J. H., and I. S. Kang, 2005: Theoretical examination of a multi-model composite for seasonal prediction. *Geophys. Res. Lett.*, **32**(18), L18707, doi: 10.1029/2005GL023513.

Yuan, H. L., X. G. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H. M. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network. *Wea. Forecasting*, **22**, 1287–1303.