


RESEARCH LETTER

Open Access



Ensemble size versus bias correction effects in subseasonal-to-seasonal (S2S) forecasts

Ji-Young Han¹, Sang-Wook Kim^{1,2}, Chang-Hyun Park² and Seok-Woo Son^{2*} 

Abstract

This study explores the ensemble size effect on subseasonal-to-seasonal (S2S) forecasts of the European Center for Medium-Range Weather Forecasts (ECMWF) model. The ensemble forecast skill and its sensitivity to the ensemble size are assessed for the troposphere and stratosphere, and compared with theoretical estimates under the perfect model assumption. The degree of skill improvement in ensemble-mean forecasts with increasing ensemble size agrees well with theoretical estimates in the troposphere. However, in the stratosphere, increasing the ensemble size does not yield as much of the skill improvement as expected. Decomposition of the mean square skill score reveals that the weak ensemble size effect in the stratosphere is primarily caused by a large unconditional bias, which exhibits no apparent decrease with increasing ensemble size. Removing such bias significantly improves the S2S forecast skill and ensemble size effect, suggesting that bias correction is crucial for S2S forecasts, especially in the stratosphere.

Keywords Subseasonal-to-seasonal (S2S) forecast, Ensemble size effect, Bias correction

Introduction

Subseasonal-to-seasonal (S2S) forecasts target time scales ranging from two weeks to two months, bridging the gap between weather and climate predictions (Robertson and Vitart 2018). Despite the importance and growing demand for S2S forecasts, the forecast skill within this time range remains lower than that of weather and climate because the S2S time scale is too long to preserve the memory of the initial conditions but is too short to be strongly influenced by the variability of the surface boundary conditions (Vitart et al. 2017).

Constructing an ensemble prediction system (EPS) is one of the straightforward and conventional methods to improve the S2S forecast skill. It is well known both theoretically and practically that increasing the number of

ensemble members leads to model forecast skill improvement, referred to as the ensemble size effect. After Leith (1974), who theoretically showed that a reasonable accuracy of Monte Carlo forecasts can be obtained with an ensemble size as small as eight, several studies have examined a reasonable ensemble size (e.g., Déqué 1997; Kumar et al. 2001). Following these studies, most operational EPSs adopt ensemble sizes ranging from 15 to 51 members by considering the balance between the computational cost and performance (Leutbecher 2019).

Previous studies on the ensemble size effect have mostly been based on the perfect model assumption, in which the model is assumed to perfectly simulate the evolution of the real atmosphere when the true state is given as an initial condition (Murphy 1988b). This assumption considers only the uncertainty in the initial conditions and ignores the model uncertainty. It is difficult to directly apply this assumption to operational forecasts because the model itself has systematic errors. Model errors play an important role in limiting the model prediction skill and predictability (Schneider et al. 2003; Stan and Kirtman 2008). Branković et al. (1990) showed

*Correspondence:

Seok-Woo Son
seokwooson@snu.ac.kr

¹ Korea Institute of Atmospheric Prediction Systems, Seoul, Republic of Korea

² School of Earth and Environmental Sciences, Seoul National University, Seoul, Republic of Korea

that systematic model errors considerably reduce the skill gain from ensemble forecasting.

This study revisits the theoretical estimate of the ensemble-mean forecast skill and compares it to the operational forecast skill derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) real-time S2S EPS, which has a large ensemble size (Vitart et al. 2017). In particular, the differences in forecast skill and ensemble size effect between the stratosphere and troposphere and those between the tropics and extratropics are explored. The reason for their differences is analyzed via skill score decomposition proposed by Murphy and Epstein (1989).

Data and methods

Data

The ECMWF real-time S2S forecast dataset is used in this study due to its large ensemble size (51 members). The ECMWF EPS simulates initial uncertainties using an ensemble of data assimilations (EDA) and singular vectors and model uncertainties using stochastic schemes (e.g., Buizza et al. 2008; Leutbecher et al. 2017). Forecasts are initialized twice a week and integrated for 46 days. All forecasts initialized from June 2015 to May 2018 (312 forecasts) are analyzed, but only the results of forecasts initialized during the December–January–February (DJF; 77 forecasts) and June–July–August (JJA; 80 forecasts) periods are presented in this study.

The ECMWF S2S prediction model uses a Variable Resolution EPS (VAREPS; Buizza et al. 2007; Vitart et al. 2008), in which the model horizontal resolution is changed during model integration. The atmospheric horizontal resolution is approximately 32 km up to forecast day 10 and approximately 64 km from day 10 onwards before the resolution upgrade in the Integrated Forecasting System (IFS) cycle 41r2 on 8 March 2016. Afterward, the resolution is set to approximately 18 km for the first 15 days and approximately 36 km thereafter. ECMWF Interim reanalysis (ERA-Interim; Dee et al. 2011) data, which are used as initial conditions, are employed as reference data for model verification and climatology calculation. In all analyses, the pentad moving-averaged data are used, and the climatology is calculated for the period 1981–2010. Verifications are conducted for 50- and 500-hPa geopotential height forecasts, which represent the stratospheric and tropospheric forecast skills, respectively, in the Northern Hemisphere extratropics (30°–90°N), tropics (30°S–30°N), and Southern Hemisphere extratropics (30°–90°S). The bias-corrected forecasts are obtained by subtracting the mean bias from the forecasts in a cross-validated way, wherein the year for which the bias is calculated is left out (Polkova et al. 2022).

Evaluation metrics

This study evaluates forecasts with the mean square skill score (MSSS), which is a diagnostic measure used in the standardized verification system for deterministic long-range forecasts (World Meteorological Organization 2006). The MSSS is defined as one minus the ratio of the mean square error (MSE) of forecasts (MSE_f) to the MSE of reference data (MSE_c):

$$MSSS(\tau, N_e) = 1 - \frac{MSE_f(\tau, N_e)}{MSE_c(\tau)}, \quad (1)$$

where

$$MSE_f(\tau, N_e) = \left\langle \left\{ \hat{f}_{ij}(\tau, N_e) - o_{ij}(\tau) \right\}^2 \right\rangle,$$

$$\hat{f}_{ij}(\tau, N_e) = \frac{1}{N_e} \sum_{k=1}^{N_e} f_{ijk}(\tau),$$

$$MSE_c(\tau) = \left\langle \left\{ c_{ij}(\tau) - o_{ij}(\tau) \right\}^2 \right\rangle.$$

Here, f is the forecast; o is the observation; c is the daily climatology of the ERA-Interim data; τ is the forecast day; and subscripts i , j , and k denote the forecast initialization time, grid point, and ensemble member, respectively. The angle bracket denotes an area-weighted average over all grid points within the region of interest (N_g) and over all initialization dates (N_f) (e.g., $MSE_f = \frac{1}{N_f \sum_{j=1}^{N_g} \cos \theta_j} \sum_{i=1}^{N_f} \sum_{j=1}^{N_g} \left\{ \hat{f}_{ij} - o_{ij} \right\}^2 \cos \theta_j$, where θ_j is the latitude in degrees). \hat{f} denotes the average of the first N_e ensemble members, where N_e ranges from 1 to 51 for the ECMWF real-time S2S forecasts. The order of ensemble members follows the numbering provided by the S2S dataset. Note that MSE_c is the climatological variance of a given variable, representing the natural variability. The MSSS equals one for perfect predictions, but typically decreases with time as forecast errors grow. When the error of the prediction system (MSE_f) is larger than the natural variability, the MSSS becomes negative, which is considered that the system loses its predictability.

Skill decomposition

Murphy and Epstein (1989) proposed a method to identify the source of the forecast skill through MSSS decomposition. The MSSS is decomposed into four terms involving the anomaly correlation coefficient (AC), conditional bias (CB), unconditional bias (UB), and difference between the mean sample and historical climatologies (CV) as follows:

$$\text{MSSS}(\tau, N_e) = \frac{\text{AC}^2(\tau, N_e) - \text{CB}^2(\tau, N_e) - \text{UB}^2(\tau, N_e) + \text{CV}^2(\tau)}{1 + \text{CV}^2(\tau)}, \quad (2)$$

where

$$\begin{aligned} \text{AC}^2(\tau, N_e) &= \frac{\left\langle \left\{ \hat{f}'_{ij}(\tau, N_e) - \langle \hat{f}'_{ij}(\tau, N_e) \rangle \right\} \left\{ o'_{ij}(\tau) - \langle o'_{ij}(\tau) \rangle \right\} \right\rangle^2}{\left\langle \left\{ \hat{f}'_{ij}(\tau, N_e) - \langle \hat{f}'_{ij}(\tau, N_e) \rangle \right\}^2 \right\rangle \left\langle \left\{ o'_{ij}(\tau) - \langle o'_{ij}(\tau) \rangle \right\}^2 \right\rangle}, \\ \text{CB}^2(\tau, N_e) &= \left\{ \text{AC}(\tau, N_e) - \frac{\sqrt{\left\langle \left\{ \hat{f}'_{ij}(\tau, N_e) - \langle \hat{f}'_{ij}(\tau, N_e) \rangle \right\}^2 \right\rangle}}{\sqrt{\left\langle \left\{ o'_{ij}(\tau) - \langle o'_{ij}(\tau) \rangle \right\}^2 \right\rangle}} \right\}^2, \\ \text{UB}^2(\tau, N_e) &= \frac{\left\langle \left\{ \hat{f}'_{ij}(\tau, N_e) - o'_{ij}(\tau) \right\} \right\rangle^2}{\left\langle \left\{ o'_{ij}(\tau) - \langle o'_{ij}(\tau) \rangle \right\}^2 \right\rangle}, \\ \text{CV}^2(\tau) &= \frac{\left\langle o'_{ij}(\tau) \right\rangle^2}{\left\langle \left\{ o'_{ij}(\tau) - \langle o'_{ij}(\tau) \rangle \right\}^2 \right\rangle}. \end{aligned}$$

Here, the prime denotes the anomaly with respect to the long-term climatology calculated from the ERA-Interim data. For instance, $f'_{ij} = f_{ij} - c_{ij}$ is the anomaly in the forecast field generated by the i -th initialization at j -th grid point. In this study, long-term climatology of reanalysis is taken to be the reference for both observations and forecasts (Murphy and Epstein 1989) to decompose the MSSS into the above four terms. Note that the third and fourth terms on right hand side of Eq. (2) (i.e., UB^2 and CV^2) become zero when observed and forecasted anomalies are defined against their respective climatologies which are calculated over the same period as the forecasts to be assessed (Goddard et al. 2013).

The term AC^2 in Eq. (2) is the square of the anomaly correlation coefficient and is considered as a measure of the potential skill that the prediction system can have when the biases are eliminated (Murphy 1988a; Murphy and Epstein 1989). This term ranges from zero to one, and the prediction system generally has a higher skill when this value is closer to one. The term CB^2 , which is the squared difference between the anomaly correlation coefficient and the ratio of the standard deviations of the forecast and observed anomalies, is a measure of the conditional bias. This term vanishes only when the slope of the regression line, in which the observed anomalies are regressed on the forecast anomalies, is equal to one. The

term UB^2 is the square of the ratio of the mean model bias to the standard deviation of the observed anomalies. This term is a measure of the unconditional bias and vanishes only for unbiased forecasts. Note that even if the model bias is small, UB^2 can be large if the variance of observation is sufficiently small. The difference between the conditional bias and unconditional bias is well illustrated in Fig. 5 of Bradley et al. (2019). The term CV^2 is the square of the ratio of the mean of the observed anomalies to the standard deviation of the observed anomalies (i.e., the inverse of the coefficient of variation of the observed anomalies) and is related to how much the observed state deviates from the climatology. For more information, see Murphy and Epstein (1989).

Theoretical ensemble size effect

Under the perfect model assumption, the forecast skill of the ensemble-mean forecast, determined by the MSE, can be expressed in terms of the ensemble size (N_e) and the average skill of individual ensemble members $\overline{\text{MSE}_f(\tau)}$ (Murphy 1988b):

$$\text{MSE}_f(\tau, N_e) = \frac{N_e + 1}{2N_e} \overline{\text{MSE}_f(\tau)}. \quad (3)$$

See the Appendix for a detailed derivation of Eq. (3). By substituting Eq. (3) into Eq. (1), the theoretical MSSS of the ensemble-mean forecast can also be expressed in terms of the ensemble size and the mean score of individual member forecasts $\overline{\text{MSSS}(\tau)}$ as follows:

$$\text{MSSS}(\tau, N_e) = 1 - \frac{N_e + 1}{2N_e} \{1 - \overline{\text{MSSS}(\tau)}\}. \quad (4)$$

The above two equations indicate an improved forecast skill with increasing ensemble size.

Figure 1a shows how the theoretical MSSS increases with the ensemble size for a given $\overline{\text{MSSS}}$. Adding the first few ensemble members dramatically improves the forecast skill of EPS, especially when the skill of individual ensemble members is lower. However, with increasing ensemble size, the ensemble size effect becomes saturated. The theoretical MSSS of the ensemble-mean forecast for an infinite ensemble size is

$$\text{MSSS}(\tau, \infty) = \frac{1}{2} \{1 + \overline{\text{MSSS}(\tau)}\}, \quad (5)$$

which is the maximum MSSS that the EPS can have. When the average MSSS of individual ensemble members, $\overline{\text{MSSS}}$, is lower than -1 , the MSSS of the

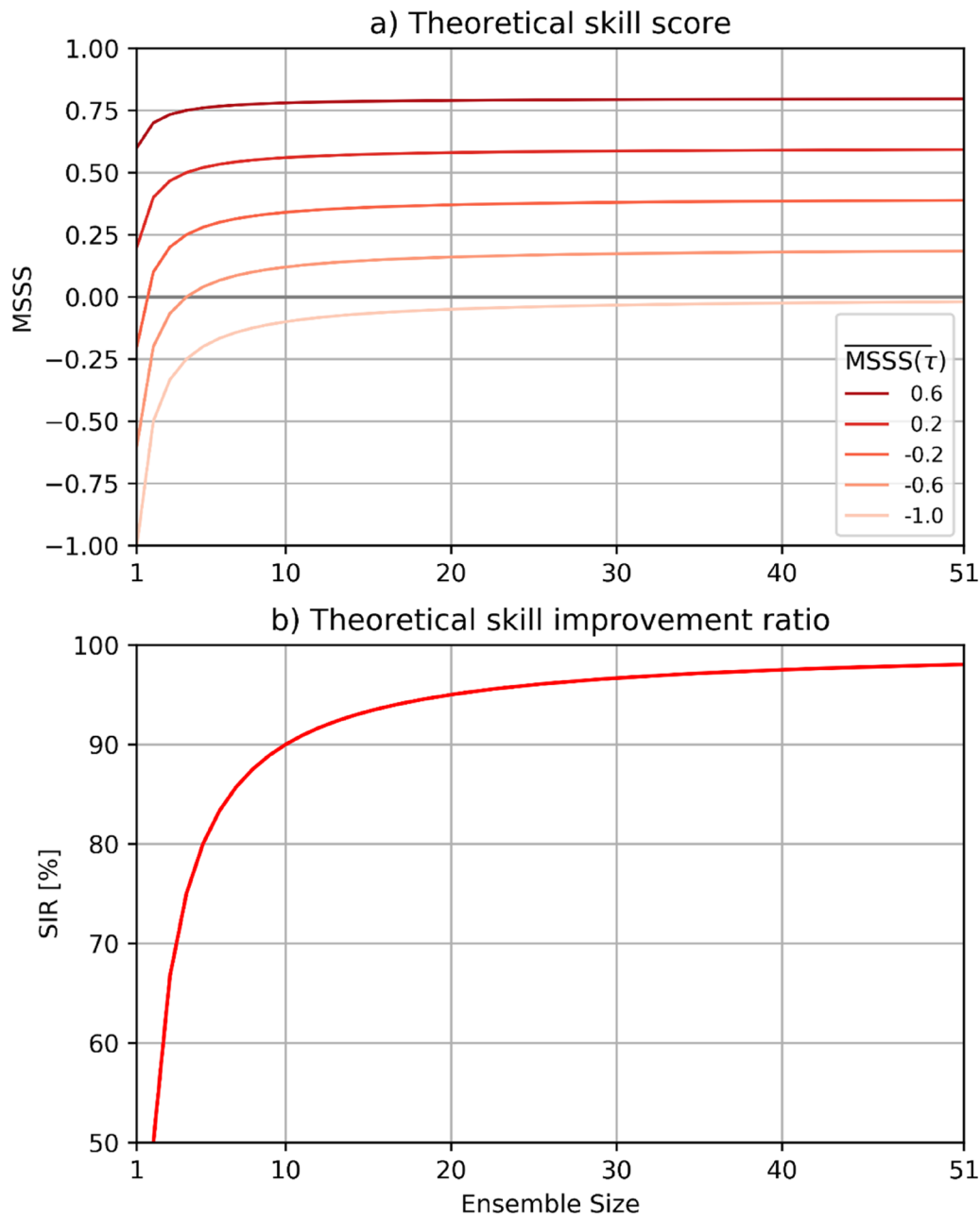


Fig. 1 **a** Theoretical MSSS of the ensemble-mean forecast for a given mean score of individual ensemble members ($\overline{\text{MSSS}}(\tau) = -1, -0.6, -0.2, 0.2$ and 0.6 ; represented by different colors) and **b** SIR multiplied by 100 (%) with varying ensemble sizes

ensemble-mean forecast becomes negative, indicating a loss of forecast skill even with an infinite ensemble size.

To measure the degree of skill improvement achieved by the ensemble-mean forecast, we define the skill improvement ratio (SIR) as the ratio of the increase in the MSSS attained using the ensemble-mean forecast to its theoretical maximum value with an infinite ensemble size:

$$\text{SIR}(N_e) = \frac{\text{MSSS}(\tau, N_e) - \overline{\text{MSSS}}(\tau)}{\text{MSSS}(\tau, \infty) - \overline{\text{MSSS}}(\tau)}. \quad (6)$$

Under the perfect model assumption, the SIR is given by $(N_e - 1)/N_e$, which is a function of the ensemble size only and independent of the skill of individual ensemble members (Fig. 1b). For example, 10 and 51 ensemble members lead to an increase in the MSSS equivalent to 90% and 98%, respectively, of its maximum value.

Again, it is clear from Fig. 1b that the ensemble size effect becomes saturated with increasing number of ensemble members.

Results

Figure 2 shows the temporal evolution of the MSSSs of individual ensemble members calculated from Eq. (1) (solid gray lines) and their mean (solid black line) as

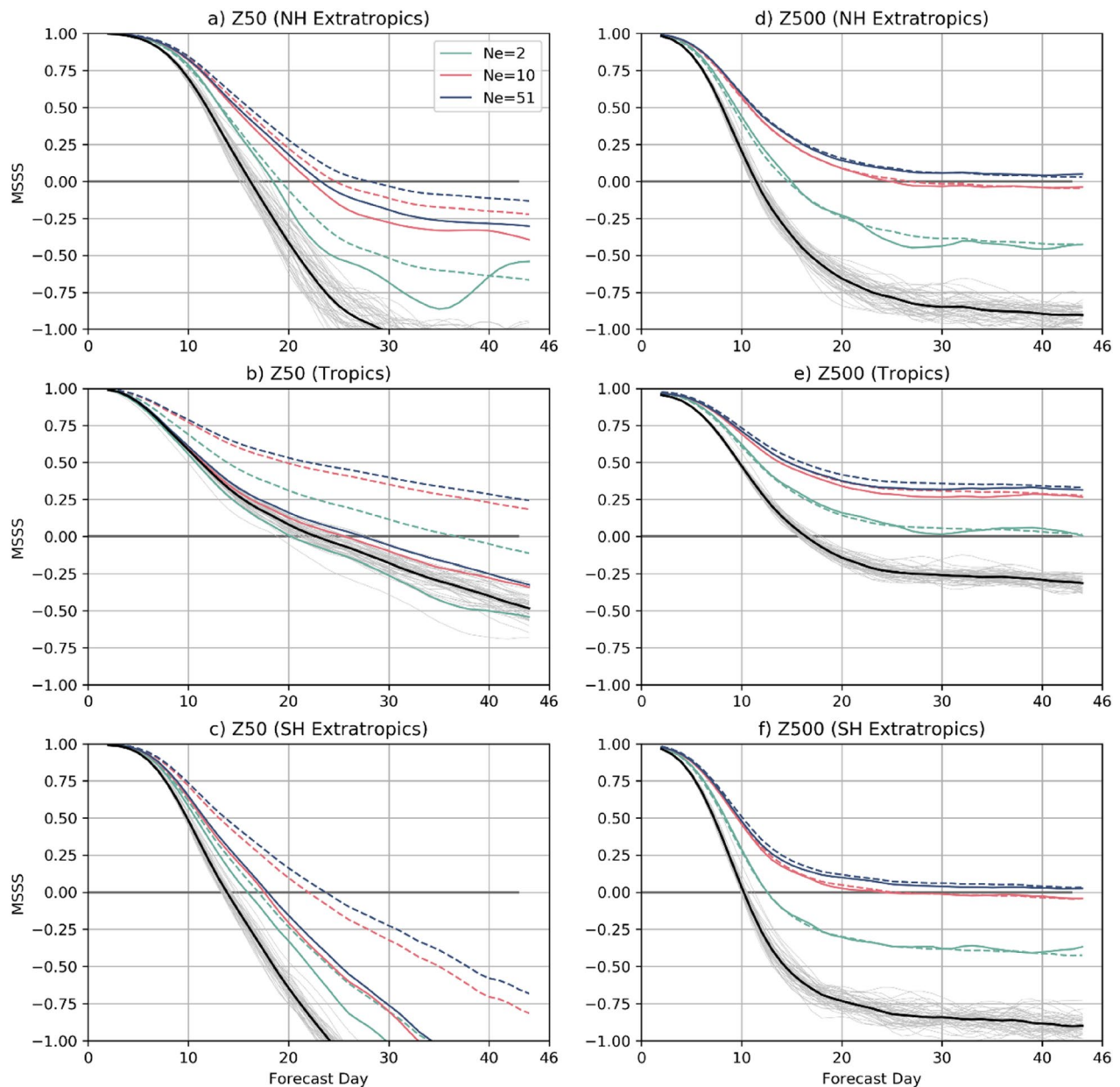


Fig. 2 Temporal evolution of the MSSSs for (left) 50- and (right) 500-hPa geopotential height forecasts initialized during the DJF 2015–2018 period in the (top) Northern Hemisphere extratropics, (middle) tropics, and (bottom) Southern Hemisphere extratropics of individual ensemble members (solid gray lines) and their mean (solid black line) as well as the MSSSs of the ensemble mean with varying ensemble sizes ($N_e=2, 10$, and 51 ; solid colored lines). The corresponding theoretical MSSSs of the ensemble-mean forecast calculated from Eq. (4) with varying ensemble sizes are shown as dashed colored lines

well as the MSSSs of the ensemble mean with varying ensemble sizes ($N_e=2, 10$, and 51 ; solid colored lines) for 50- and 500-hPa geopotential height forecasts initialized during the DJF 2015–2018 period in the Northern Hemisphere extratropics, tropics, and Southern Hemisphere extratropics, and the corresponding theoretical MSSSs of the ensemble-mean forecast calculated from Eq. (4) (dashed colored lines). At both the 50- and 500-hPa levels in all regions, a high degree of forecast skill is maintained over the first few forecast days, but the skill sharply decreases afterward. The stratospheric forecast skill decreases slower and, therefore, remains higher than the tropospheric skill on earlier forecast days. However, it decreases continuously with time during the integration period, in contrast to the tropospheric forecast skill which becomes nearly constant after approximately 30 forecast days. The stratospheric MSSSs eventually become much lower than the tropospheric ones on later forecast days (solid colored lines in Fig. 2). The forecast skill in the tropics is higher than that in the extratropics, which can be partly explained by the strong modulation of the tropical atmosphere by the underlying sea surface temperature (Shukla 1998).

While the operational forecast skill is in good agreement with the theoretical estimate in the troposphere (compare the solid and dashed colored lines in the right column), there is a large discrepancy between them in the stratosphere (left column). The stratospheric MSSSs are substantially lower than the corresponding theoretical values except after 40 forecast days with $N_e=2$ in the Northern Hemisphere extratropics (Fig. 2a). This is particularly true in the tropics and summer hemisphere (Fig. 2b and c). The tropical stratospheric skill of the ensemble-mean forecast with $N_e=2$ is even lower than the average skill of individual ensemble members (compare the solid dark cyan and black lines in Fig. 2b). The skill improvement is also relatively minor in the tropics and summer hemisphere when the ensemble size is increased from $N_e=2$ to $N_e=51$ (compare the solid dark cyan and dark blue lines in Fig. 2b and c). Similar results are also obtained with the JJA forecasts (Additional file 1: Fig. S1). These results indicate that the ensemble size effect is different depending on the region.

The discrepancy between the estimated and operational forecast skills implies that increasing the ensemble size does not necessarily ensure as much of an improvement in forecast skill as expected. This also indicates that the perfect model assumption does not hold for operational forecasts, especially in the stratosphere. In this study, the source of the weak ensemble size effect in the stratosphere is identified with the skill decomposition method described in the Data and Method section.

Figures 3 and 4 show the results of MSSS decomposition, i.e., AC^2 , CB^2 , UB^2 , and CV^2 , for the 50- and 500-hPa geopotential height forecasts initialized during the boreal winter in the Northern Hemisphere extratropics and tropics, respectively. The overall features of the Southern Hemisphere extratropics are similar to those of the Northern Hemisphere extratropics and are therefore not presented here. AC^2 decreases more slowly with the forecast days in the stratosphere than in the troposphere. Therefore, its values in the stratosphere are higher than those in the troposphere (first row in Figs. 3 and 4). In particular, in the tropical stratosphere, the forecasts maintain relatively high AC^2 values during the forecast period. This implies that the model possesses a higher potential forecast skill in the stratosphere than in the troposphere, especially in the tropics. AC^2 increases with increasing ensemble size, as shown in previous studies (e.g., Leith 1974; Murphy 1988b; Branković et al. 1990), in both the stratosphere and troposphere.

The CB^2 tends to increase with the forecast days, but the increase rate decreases and even becomes negative on later forecast days (second row in Figs. 3 and 4). In the extratropics, its values in the stratosphere are lower than those in the troposphere on earlier forecast days but higher on later forecast days (Fig. 3b and f). However, in the tropics, CB^2 is much smaller in the stratosphere than that in the troposphere throughout the forecast period (Fig. 4b and f). The conditional bias, which represents the linear slope between the forecast and observed states, can be understood as errors in the amplitude of eddies deviating from the mean state. Son et al. (2020) who examined the extratropical prediction skill of the S2S prediction models showed that eddy amplitude error is higher in the extratropical troposphere than in the stratosphere on earlier forecast days, and this may be attributable to the failure to predict the amplitude of synoptic scale eddies in the troposphere. CB^2 effectively decreases with increasing ensemble size. Even with 10 ensemble members ($N_e=10$), CB^2 becomes close to zero in the tropics.

While both AC^2 and CB^2 make the MSSS increase with the ensemble size, UB^2 exhibits no obvious decrease with increasing ensemble size (Fig. 4c). Note that UB^2 in the tropical stratosphere with $N_e=2$ is larger than that of individual forecasts. Hence, UB^2 inhibits the ensemble size effect. In the stratosphere, UB^2 increases more dramatically with the forecast days in the tropics than in the extratropics (Figs. 3c and 4c). In the troposphere where the model bias is small (Lawrence et al. 2022), UB^2 is close to zero throughout the forecast period (Figs. 3g and 4g). Note that despite the large model bias related to the wintertime polar vortices (Lawrence et al. 2022), UB^2 in

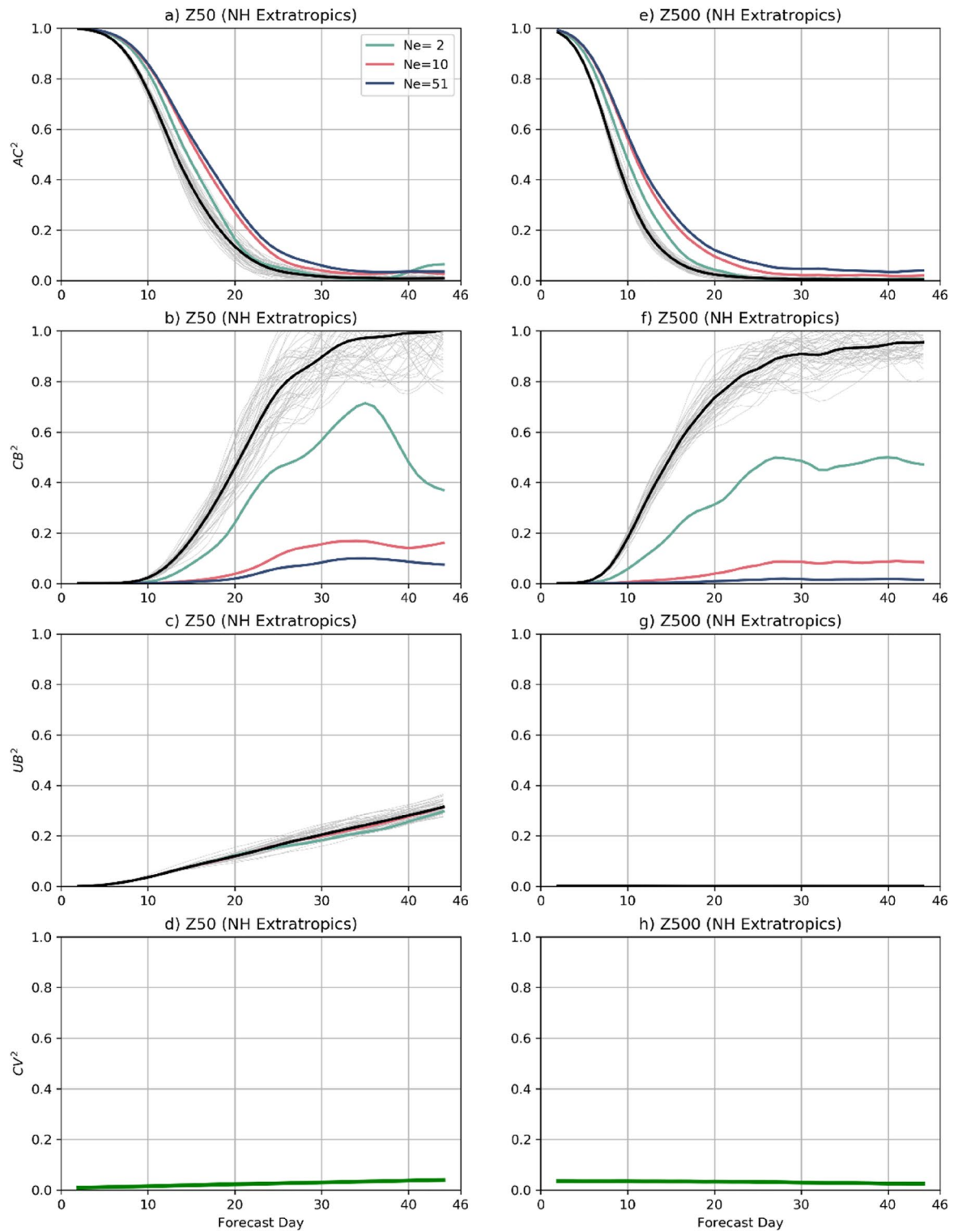


Fig. 3 Temporal evolution of the four terms (from top to bottom: AC^2 , CB^2 , UB^2 , and CV^2) of MSSS decomposition for (left) 50- and (right) 500-hPa geopotential height forecasts initialized during the DJF 2015–2018 period in the Northern Hemisphere extratropics of individual ensemble members (gray lines) and their mean (black line) as well as those of the ensemble mean with varying ensemble sizes ($N_e = 2, 10$, and 51 ; colored lines). Note that CV^2 , which is independent of the forecasts, is indicated with a green line

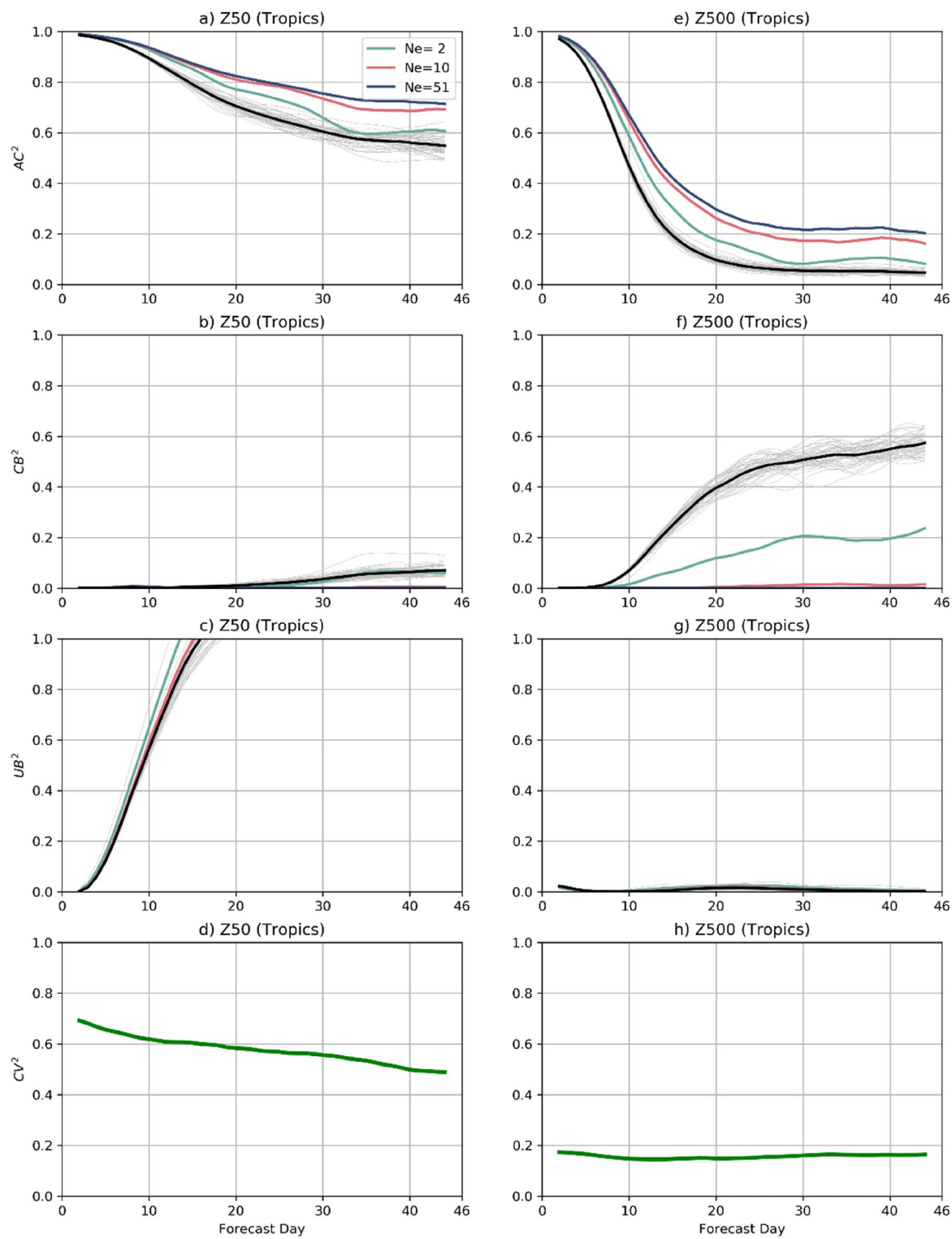


Fig. 4 Same as Fig. 3, but in the tropics

the extratropical stratosphere is relatively small because the variance of observation is large there (Additional file 1: Fig. S2). In contrast, in the tropical stratosphere where the variance of observation is small, a considerably large UB^2 is observed. Note that UB can be understood as a failure to predict the mean state. Son et al. (2020) showed that the model errors associated with the zonal-mean flow grow rapidly in the stratosphere. The tropical stratospheric bias might be related to the quasi-biennial oscillation (QBO) which dominates the circulation of the

tropical stratosphere (Lawrence et al. 2022). The CV^2 , which is independent of the forecasts, is large in the tropics because of a small variance of observation, while it is negligible in the extratropics.

Because UB^2 inhibits the ensemble size effect, one of the simplest ways to enlarge the gain obtained from ensemble forecasts is to subtract the model mean bias from the forecast, referred to as bias correction. Figure 5 shows the temporal evolution of the MSSs of the forecasts after bias correction. Compared with the results

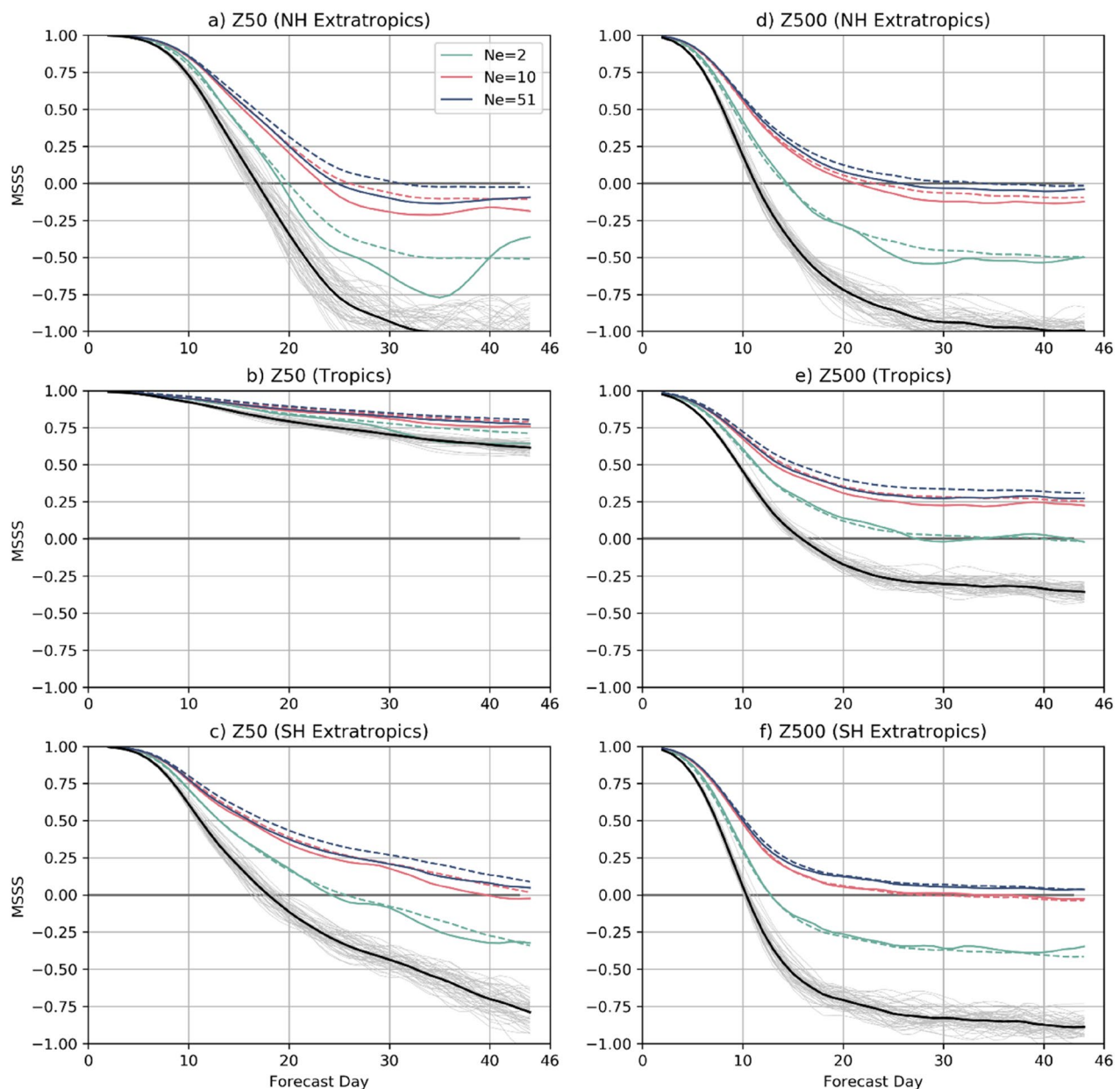


Fig. 5 Same as Fig. 2, but after bias correction

in Fig. 2, it is apparent that the MSSS increases considerably in the stratosphere. In particular, in the tropical stratosphere, even the individual forecasts after bias correction maintain MSSSs greater than 0.5 until the end of the forecast period. Note that lower degrees of freedom could be the physical reason why the geopotential height forecasts in the tropical stratosphere requires a smaller ensemble size than other regions to reach the saturation of skills. More importantly, the discrepancy between the estimated and operational skills in the stratosphere substantially decreases. As shown in Additional file 1: Fig. S3, the operational skill after bias correction becomes almost the same as the theoretical skill during the JJA period. Therefore, the expected forecast skill improvement with increasing ensemble size is attained after bias correction even in the stratosphere. Similar results are also obtained in temperature forecasts as shown in Additional file 1: Figs. S4 and S5. These changes in the operational forecast skill and ensemble size effect are more pronounced at the S2S time scale.

Here, it should be stated that the above analyses aim not to evaluate the skill of the real-time forecasts, but to quantify the importance of the unconditional bias and its relationship to the theoretical estimate. For this purpose, the model mean bias is corrected with the real-time forecasts (see Data and Methods). However, in the operational forecasts, the bias correction is conducted with the long-term reforecasts. By considering the application of this study to the operational forecasts, the same analyses are repeated by correcting the model mean bias with the long-term reforecasts over the past 20 years which are available from the S2S prediction project website. Although there is a difference in the degree to which skill is improved, similar results are obtained (Additional file 1: Fig. S6). This result indicates that the ensemble size effect closer to the theoretical estimate can be obtained by removing the unconditional bias regardless of the details of the bias correction method.

Summary

This study explores the forecast skill and its dependency on the ensemble size in ECMWF real-time S2S forecasts. The results are particularly compared with theoretical estimates obtained under the perfect model assumption. In the troposphere, a high degree of skill is maintained for the first few days, after which the forecast skill rapidly declines and then stabilizes, and exhibit a good agreement with the theoretical estimate. The forecast skill in

the stratosphere is higher than that in the troposphere on earlier forecast days, but decreases continuously, eventually becoming lower than the tropospheric skill on later forecast days. The stratospheric skill decreases much faster than the theoretical estimate, especially in the tropics.

The discrepancy between the estimated and operational skills in the stratosphere is mostly attributed to the unconditional bias resulting from model drift. By applying bias correction, the stratospheric forecast skill is substantially increased, becoming comparable to the theoretical estimate. The skill improvement with increasing ensemble size also follows the theoretical estimate. This result does not mean that the model can be assumed to be perfect, but should be understood in the limited sense that it is possible to expect an ensemble size effect obtained under the perfect model assumption when the unconditional bias is negligible. The results obtained in this study are not unique to the ECMWF model. Although not shown, essentially the same results have been found in analyses of the Centre National de Recherches Météorologiques (CNRM) model of Météo-France.

While bias correction can improve the ensemble size effect, the extent to which it can do so practically depends on various factors such as the specific bias correction method, the quality of reforecasts or climatology used for bias correction, and the target variable. The importance of bias correction for operational forecasts may not be surprising (e.g., Son et al. 2020). However, this study quantifies its importance and relationship to theoretical estimates. The optimal ensemble size for operational S2S forecasts is also estimated. For example, it is revealed that an ensemble size of 10 to 20 yields a forecast skill that is approximately 90% to 95% of the theoretical maximum skill.

Appendix

A detailed derivation of Eq. (3) is as follows:

$$\begin{aligned} \text{MSE}_f(\tau, N_e) &= \left\langle \left\{ \hat{f}_{ij}(\tau, N_e) - o_{ij}(\tau) \right\}^2 \right\rangle \\ &= \left\langle \left\{ \hat{f}_{ij}(\tau, N_e) - \bar{f}_{ij}(\tau) \right\}^2 + \left\{ o_{ij}(\tau) - \bar{f}_{ij}(\tau) \right\}^2 \right. \\ &\quad \left. - 2 \left\{ \hat{f}_{ij}(\tau, N_e) - \bar{f}_{ij}(\tau) \right\} \left\{ o_{ij}(\tau) - \bar{f}_{ij}(\tau) \right\} \right\rangle \end{aligned}$$

$$= \left\langle \frac{1}{N_e^2} \sum_{k=1}^{N_e} \left\{ f_{ijk}(\tau) - \bar{f}_{ij}(\tau) \right\}^2 + \frac{2}{N_e^2} \sum_{k=1}^{N_e} \sum_{l=k+1}^{N_e} \left\{ f_{ijk}(\tau) - \bar{f}_{ij}(\tau) \right\} \left\{ f_{ijl}(\tau) - \bar{f}_{ij}(\tau) \right\} \right. \\ \left. + \left\{ o_{ij}(\tau) - \bar{f}_{ij}(\tau) \right\}^2 - 2 \left\{ \hat{f}_{ij}(\tau, N_e) - \bar{f}_{ij}(\tau) \right\} \left\{ o_{ij}(\tau) - \bar{f}_{ij}(\tau) \right\} \right\rangle,$$

where an overbar denotes an average over a forecast probability density function (p.d.f), and the second and last terms, i.e., averages of the sum of random covariances, are zero. Then, by denoting the variance of a forecast p.d.f by D ,

$$\text{MSE}_f(\tau, N_e) = \frac{1}{N_e} \langle D \rangle + \langle d_o \rangle \approx \{(1 + N_e)/N_e\} \langle D \rangle,$$

where $d_o = \left\{ o_{ij}(\tau) - \bar{f}_{ij}(\tau) \right\}^2$, and $\langle d_o \rangle = \langle D \rangle$ is assumed for simplicity. Thus, using that $\text{MSE}_f(\tau, 1) = 2\langle D \rangle$ for $N_e = 1$, we obtain the following equation indicating that an ensemble mean forecast is superior to an individual forecast:

$$\text{MSE}_f(\tau, N_e) = \{(1 + N_e)/2N_e\} \text{MSE}_f(\tau, 1),$$

which corresponds to Eq. (5) in Murphy (1988b).

Abbreviations

AC	Anomaly correlation coefficient
CB	Conditional bias
CNRM	Centre National de Recherches Météorologiques
CV	Difference between the mean sample and historical climatology
DJF	December–January–February
ECMWF	European Center for Medium-Range Weather Forecasts
EPS	Ensemble prediction system
ERA-Interim	ECMWF Interim reanalysis
IFS	Integrated Forecasting System
JJA	June–July–August
MSE	Mean square error
MSSS	Mean square skill score
QBO	Quasi-Biennial oscillation
S2S	Subseasonal-to-seasonal
SIR	Skill improvement ratio
UB	Unconditional bias
VAREPS	Variable resolution EPS

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40562-023-00292-9>.

Additional file 1: **Fig. S1** Same as Fig. 2, but for the JJA forecasts. **Fig. S2** Global distribution of the standard deviation of observation for (a) 50- and (b) 500-hPa geopotential heights. **Fig. S3** Same as Fig. 5, but for the JJA forecasts. **Fig. S4** Same as Fig. 2, but for temperature forecasts. **Fig. S5** Same as Fig. 5, but for temperature forecasts. **Fig. S6** Same as Fig. 5, but with the long-term reforecasts.

Author contributions

JYH organized, wrote, and revised the manuscript. SWK collected and analyzed data and prepared the first draft of the manuscript. CHP contributed

to data collection and analysis. SWS performed supervision and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was carried out through the R&D project “Development of a Next-Generation Numerical Weather Prediction Model by the Korea Institute of Atmospheric Prediction Systems (KIAPS)”, funded by the Korea Meteorological Administration (KMA2020-02212). This work was also supported by Korea Environment Industry and Technology Institute (KEITI) through “Climate Change R&D Project for New Climate Regime”, funded by Korea Ministry of Environment (MOE) (2022003560004).

Availability of data and materials

The ECMWF S2S real-time forecast data (<http://apps.ecmwf.int/datasets/data/s2s>) and ERA-Interim data (https://apps.ecmwf.int/datasets/data/interim_full_daily) are available online.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 6 December 2022 Accepted: 4 August 2023

Published online: 19 August 2023

References

- Bradley AA, Demargne J, Franz KJ (2019) Attributes of forecast quality. In Handbook of hydrometeorological ensemble forecasting Springer, Berlin
- Branković Č, Palmer TN, Molteni F, Tibaldi S, Cubasch U (1990) Extended-range predictions with ECMWF models: time-lagged ensemble forecasting. Q J R Meteorol Soc 116:867–912
- Buizza R, Bidlot J-R, Wedi N, Fuentes M, Hamrud M, Holt G, Vitart F (2007) The new ECMWF VAREPS (variable resolution ensemble prediction system). Q J R Meteorol Soc 133:681–695
- Buizza R, Leutbecher M, Isaksen L (2008) Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. Q J R Meteorol Soc 134:2051–2066
- Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beijaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavaloto C, Thépaut J-N, Vitart F (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597
- Déqué M (1997) Ensemble size for numerical seasonal forecasts. Tellus 49A:74–86
- Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Khari V, Merryfield W, Deser C, Mason SJ, Kirtman BP, Msadek R, Sutton R, Hawkins E, Fricker T, Hegerl G, Ferro CAT, Stephenson DB, Meecl GA, Stockdale T, Burgman R, Greene AM, Kushnir Y, Newman M, Carton J, Fukumori I, Delworth T (2013) A verification framework for interannual-to-decadal predictions experiments. Clim Dyn 40:235–272
- Kumar A, Barnston AG, Hoerling MP (2001) Seasonal predictions, probabilistic verifications, and ensemble size. J Clim 14:1671–1676
- Lawrence ZD, Abalos M, Ayazagüena B, Barriopedro D, Butler AH, Calvo N, de la Cámara A, Chalton-Perez A, Domeisen DIV, Dunn-Sigouin E,

- Carcía-Serrano J, Garfinkel CI, Hindley NP, Jia L, Jucker M, Karpechko AY, Kim H, Lang AL, Lee SH, Lin P, Osman M, Palmeiro FM, Perlwitz J, Polichtchouk I, Richter JH, Schwartz C, Son S-W, Statnaia I, Taguchi M, Tyrrell NL, Wright CJ, Wu RW-Y (2022) Quantifying stratosphere biases and identifying their potential sources in subseasonal forecast systems. *Weather and Climate Dynamics* 3:977–1001
- Leith CE (1974) Theoretical skill of Monte Carlo forecasts. *Mon Weather Rev* 102:409–418
- Leutbecher M (2019) Ensemble size: How suboptimal is less than infinity? *Q J R Meteorol Soc* 145:107–128
- Leutbecher M, Lock S-J, Ollinaho P, Lang STK, Balsamo G, Bechtold P, Bonavita M, Christensen HM, Diamantakis M, Dutra E, English S, Fisher M, Forbes RM, Goddard J, Haiden T, Hogan RJ, Juricke S, Lawrence H, MacLeod D, Magnusson L, Malardel S, Massart S, Sandu I, Smolarkiewicz PK, Subramanian A, Vitart F, Wedi N, Weisheimer A (2017) Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Q J R Meteorol Soc* 143:2315–2339
- Murphy AH (1988a) Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon Weather Rev* 116:2417–2424
- Murphy JM (1988b) The impact of ensemble forecasts on predictability. *Q J R Meteorol Soc* 114:463–493
- Murphy AH, Epstein ES (1989) Skill scores and correlation coefficients in model verification. *Mon Weather Rev* 117:572–581
- Polkova I, Schaffer L, Aarnes Ø, Baehr J (2022) Seasonal climate predictions for marine risk assessment in the Barents sea. *Climate Services* 26:100291
- Roberts CD, Senan R, Molteni F, Boussetta S, Mayer M, Keely SPE (2018) Climate model configurations of the ECMWF integrated forecasting system (ECMWF-IFS cycle 43r1) for HighResMIP. *Geosci Model Dev* 11:368–372
- Robertson AW, Vitart F (2018) Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting. Elsevier, Amsterdam
- Schneider EK, Dewitt D, Rosati A, Kirtman BP, Ji L, Tribbia JJ (2003) Retrospective ENSO forecasts: sensitivity to atmospheric model and ocean resolution. *Mon Weather Rev* 131:3038–3060
- Shukla J (1998) Predictability in the midst of chaos: a scientific basis for climate forecasting. *Sicence* 282:728–731
- Son S-W, Kim H, Song K, Kim S-W, Martineau P, Hyun Y-K, Kim Y (2020) Extratropical prediction skill of the subseasonal-to-seasonal (S2S) prediction models. *J Geophys Res Atmos* 125:e2019JD031273
- Stan C, Kirtman BP (2008) The influence of atmospheric noise and uncertainty in ocean initial conditions on the limit of predictability in a coupled GCM. *J Clim* 21:3487–3503
- Vitart F, Buizza R, Balmaseda MA, Balsamo G, Bidlot J-R, Bonet A, Fuentes M, Hofstadler A, Molteni F, Palmer TN (2008) The new VarEPS-monthly forecasting system: a first step towards seamless prediction. *Q J R Meteorol Soc* 134:1789–1799
- Vitart F, Ardilouze C, Bonet A, Brookshaw A, Chen M, Codorean C, Déqué M, Ferranti L, Fucile E, Fuentes M, Hendon H, Hodgson J, Kang H-S, Kumar A, Lin H, Liu G, Liu X, Malguzzi P, Mallas I, Manoussakis M, Mastrangelo D, MacLachlan C, McLean P, Minami A, Mladek R, Nakazawa T, Najm S, Nie Y, Rixen M, Robertson AW, Ruti P, Sun C, Takaya Y, Tolstykh M, Venuti F, Waliser D, Woolnough S, Wu T, Won D-J, Xiao H, Zaripov R, Zhang L (2017) The subseasonal to seasonal (S2S) prediction project database. *Bull Am Meteorol Soc* 98:163–173
- World Meteorological Organization (WMO) (2006) Standardized verification system (SVS) for long-range forecasts (LRF) (attachment II.8) in the manual on the global data-processing and forecasting system (WMO-No. 485)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)