이학석사 학위논문

Optimal ensemble size for Sub-seasonal to Seasonal (S2S) prediction system

계절내 및 계절 (S2S) 예측을 위한 적절한 앙상블 크기 설정에 대한 연구

2019 년 8 월

서울대학교 대학원

지구환경과학부

김 상 욱

Optimal ensemble size for Sub-seasonal to Seasonal (S2S) prediction system

계절내 및 계절 (S2S) 예측을 위한 적절한 앙상블 크기 설정에 대한 연구

지도교수 손 석 우

이 논문을 이학석사 학위논문으로 제출함
2019 년 6 월

서울대학교 대학원
지구환경과학부
김 상 욱

김상욱의 이학석사 학위논문을 인준함
2019 년 6 월

위 원 장 ______________________ (인)

부위원장 ______________________ (인)

위 원 ______________________ (인)

Abstract

In this study, the optimal ensemble size and the factor affecting its determination for Sub-seasonal to Seasonal (S2S) prediction are explored. The results show that the prediction skill, which is quantified by the mean square skill score (MSSS), increases with increasing ensemble size in a perfect model with an unbiased and reliable ensemble spread to observed variance. In this idealized prediction system, 10 to 20 ensemble members lead to 90% to 95% of the skill improvement in the theoretical maximum (infinite ensemble size). This theoretical estimation is applied to the European Centre for Medium-Range Weather Forecasts (ECMWF) S2S real-time forecast that consists of 51 ensemble members. The MSSS of the 500-hPa geopotential height, which represents tropospheric skill, is in good agreement with the theoretical estimation, indicating that approximately 10 to 20 members are needed to obtain a skill improvement close to the theoretical maximum. However, the stratospheric skill, verified at the 50-hPa geopotential height, is substantially lower than the theoretical estimation. This result is particularly true in the tropical stratosphere, where only 20-30% skill improvement of the theoretical maximum can be obtained even when using all ensemble members. The substantial overestimation of the stratospheric skill is mainly due to model mean bias. The role of mean bias in the ensemble size effect is highly dependent on ensemble spread and natural variability. By removing the bias, overestimation in the stratosphere can be reduced so that it is possible for 10-20 members to be optimal. Extending the results to all the levels and seasons available in the S2S dataset indicates that the effect of mean bias is remarkable in the tropical and summer hemispheric upper stratosphere.

Keyword: Sub-seasonal to Seasonal (S2S) prediction, Ensemble prediction system (EPS), Ensemble size effect, Optimal ensemble size, Prediction skill
Student Number: 2017−20309

Table of Contents

1. Introduction

Sub-seasonal to seasonal (S2S) prediction targets a forecast from 2 weeks to 2 months and serves as a bridge between weather and climate forecast (e.g., Robertson and Vitart, 2018). Prediction of this time-scale is valuable information for decision-makers since various social and economic decisions fall into this time range, such as disaster mitigation, public health care, and the economy (Vitart et al., 2017). As the importance of this time-scale prediction is taken into account, the World Meteorological Organization (WMO) launched the project for the S2S forecast to constitute collaborative research among the leading group around the world (Vitart et al., 2012; WMO, 2013).

Improving the S2S predictability, which has been considered a "predictability desert," is one of the main issues of this research project. Several studies tried to verify and improve S2S predictability (e.g., Vitart, 2017; Lim et al., 2018). These studies suggested that the "predictability desert" occurs because the S2S time-scale is too long to preserve the memory of an initial condition, and at the same time, it is too short to be influenced by a boundary condition that has a long memory.

Constructing an ensemble forecast system is one of the straightforward and conventional methods to enhance the predictability of the forecast system (Gneiting and Raftery, 2005). It is well known, theoretically and practically, that increasing the number of ensemble members leads to improved model predictability. However, since expanding ensemble size demands expensive computational resources, determining the optimal ensemble size in operating an ensemble system is an essential concern for establishing an efficient forecast system. Previous studies showed that economically efficient predictions could be made for 10-20 ensemble members for probabilistic forecasts (Kumar et al., 2001; Richardson, 2001).

While optimal ensemble size for probabilistic forecasts has been discussed, few works have analyzed for deterministic forecasts (Deque, 1997). It is probably due to a deterministic forecast using an ensemble mean has not been considered the primary purpose of establishing an ensemble prediction system, although it is still a useful technique in monthly and seasonal prediction. Partly, Murphy (1988b) and Brankovic et al. (1990) proved that the deterministic model error decreases as the number of ensembles increases.

To investigate how the predictability of a forecast system can be improved, a perfect model, which completely predicts the evolution of real atmosphere when given a true state as an initial condition, should be assumed (Leith, 1974; Murphy 1988). This assumption implies that the model uncertainty is ignored and that only the uncertainty in an initial condition is considered. However, it is hard to apply this assumption directly to a practical forecast system. It is particularly true for S2S prediction systems since the models usually have its own bias to the specific direction which increases with the forecast step. In other words, in a practical forecast system, expanding ensemble size does not always guarantee an improvement in the prediction skill, as referred to in previous studies (e.g., Murphy 1988b; Brankovic et al., 1990). Since this bias is inhomogeneous in region and season, the skill improvement through expanding ensemble size probably depends on the verifying region and season.

In this vein, this study revisits the theoretical discussion by Murphy (1988b) and applies it to the practical model results, the European Centre for Medium-range Weather Forecast (ECMWF) S2S real-time forecast data, which consists of an adequate number of ensemble members of 51. The difference in the predictability and the ensemble size effect by region and season will be explored, and the reason for this difference will be partly discussed.

In chapter 2, the optimal ensemble size for the idealized forecast system is discussed. It is applied to the ECMWF model in chapter 3, and the possible causes of the discrepancy between the estimated and practical skills are also described. Finally, chapter 4 summarizes the overall results.

2. Methodology

2.1. Error Estimation in a Perfect Model

One of the most basic methods for quantitatively verifying a forecast system is the mean square error (MSE), which is calculated based on the error between the forecast and the corresponding observation. The MSE indicates the Euclidian distance between them in terms of geometry. From this point of view, the MSE for a certain forecast step of τ and e ensemble members can be defined as

$$MSE_f(\tau,\ e) = \frac{1}{N_f}\frac{1}{N_g}\sum_{i=1}^{N_f}\sum_{j=1}^{N_g}\left(\hat{f}_{ij}(\tau,e) - o_{ij}(\tau)\right)^2\ w_j \qquad (1)$$

$$where\ \ \hat{f}_{ij}(\tau,e) = \frac{1}{e}\sum_{k=1}^{e} f_{ij}(\tau,k).$$

f_{ij} represents the forecast at initialization time i and grid point j, and o_{ij} indicates the corresponding observation. $\hat{f}$ is the ensemble result with the e members containing a control member. w_j represents latitudinal weighting, and N_f and N_g indicate the total initialization time and the number of grid points, respectively. The MSE will be 0 when the forecast system completely predicts the observation (the perfect prediction) and will increase infinitely as the distance between them increases.

It is possible to describe the change in the MSE as a function of the ensemble size and the MSE of the individual ensemble member (Murphy, 1988b). As mentioned in the introduction, a perfect model that can fully predict the atmospheric condition when the true state is given as the initial condition should be assumed. Under this perfect model condition, the forecast system shares statistical characteristics, such as the mean and variance, with the observation. Moreover, the nonlinearity of a forecast system is considered small enough, so that all ensemble members from random perturbed initial conditions can be considered reliable. In this condition, the true state (or the observed state) can be thought of as one of the elements extracted from the probability density function (PDF) of the infinite forecast, and the e ensemble members are defined by the same number of samples from the PDF. When $\bar{f}$ is the infinite member ensemble mean, the MSE of the forecast, initialized in i, at forecast step τ with ensemble number e, $MSE_{f_i}(\tau, e)$, can be written as follows ($\langle\ \rangle$ represents an average over the domain).

$$MSE_{f_i}(\tau, e) = \frac{1}{N_g}\sum_{j=1}^{N_g}\left(\hat{f}_{ij}(\tau, e) - o_{ij}(\tau)\right)^2 w_j$$

$$= \langle\left(\hat{f}_{ij} - o_{ij}\right)^2\rangle = \langle\left(\left(\hat{f}_{ij} - \bar{f}_{ij}\right) - \left(o_{ij} - \bar{f}_{ij}\right)\right)^2\rangle$$

$$= \langle\left(\hat{f}_{ij} - \bar{f}_{ij}\right)^2 + \left(o_{ij} - \bar{f}_{ij}\right)^2 - 2\left(\hat{f}_{ij} - \bar{f}_{ij}\right)\left(o_{ij} - \bar{f}_{ij}\right)\rangle$$

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

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

In this case, if the finite ensemble extracted from the forecast PDF is sufficiently reliable, and averaged over infinite repetitive extraction the last term is approximated to zero. When the variance in PDF is represented by D, the estimated MSE for initialization i ($\overline{MSE_{f_i}}(\tau, e)$) is

$$\overline{MSE_{f_i}}(\tau, e) = \langle \frac{1}{e} \left(\frac{1}{e} \sum_{k=1}^{e} \left(f_{ij}(\tau, k) - \bar{f}_{ij} \right)^2 \right) + \left(o_{ij} - \bar{f}_{ij} \right)^2 \rangle$$

$$= \langle \frac{D}{e} + \left(o_{ij} - \bar{f}_{ij} \right)^2 \rangle = \frac{\langle D \rangle}{e} + \langle \left(o_{ij} - \bar{f}_{ij} \right)^2 \rangle.$$

When the above equation is averaged over large enough forecast events, the last term on the right-hand side would be equal to $\langle D \rangle$ since the true state is sampled from the same PDF with forecasts. Therefore, the estimated MSE containing ensemble size e ($\overline{MSE_f}(\tau, e)$) and the minimum (when using an infinite ensemble size) are obtained as follows.

$$\overline{MSE_f}(\tau, e) = \frac{1}{N_f} \sum_{i=1}^{N_f} \overline{MSE_{f_i}}(\tau, e) = \frac{\langle D \rangle}{e} + \langle D \rangle$$

$$\widetilde{MSE_f}(\tau) \approx \overline{MSE_f}(\tau, 1) = 2\langle D \rangle$$

$$\overline{MSE_f}(\tau, e) = \frac{e+1}{2e}\overline{MSE_f}(\tau, 1) \approx \frac{e+1}{2e}\widetilde{MSE_f}(\tau) \tag{2}$$

$$\overline{MSE_f}(\tau, \infty) = \frac{1}{2}\overline{MSE_f}(\tau, 1) \approx \frac{1}{2}\widetilde{MSE_f}(\tau) \tag{3}$$

Here, $\widetilde{MSE_f}(\tau)$ indicates the average MSE of individual ensemble members. Those equations imply that even the ensemble system providing an infinite number of members, cannot reduce the error to zero but by half. For more details, see Murphy (1988b).

2.2. Evaluation Metrics (MSSS and SIR)

The mean square skill score (MSSS, or skill score) is a measure of the predictability of the forecast, based on the MSE. The MSSS is defined as follows (Murphy, 1988a; Murphy and Epstein, 1989).

$$MSSS(\tau, e) = \frac{MSE_f - MSE_{ref}}{MSE_{prefect} - MSE_{ref}} = \frac{MSE_c(\tau) - MSE_f(\tau, e)}{MSE_c(\tau)} \tag{4}$$

$$where\ \ MSE_c(\tau,\ e) = \frac{1}{N_f}\frac{1}{N_g}\sum_{i=1}^{N_f}\sum_{j=1}^{N_g}\left(c_{ij}(\tau, e) - o_{ij}(\tau)\right)^2\ w_j$$

The climatological forecast is traditionally used as the reference forecast so that the MSE of the reference forecast MSE_{ref} is the same as the variance in nature. The MSE of the perfect model

$MSE_{perfect}$ is zero. c_{ij} is the climatology corresponding to the initialization date and forecast step. The skill score would be 1 for a perfect prediction and decreases toward negative infinity as the error of the forecast increases. If MSSS is lower than 0, the predictability of the forecast is lower than the reference. It can also be said that the forecast error is larger than the natural change since MSE_c indicates the variance in the observation. Therefore, it is relevant to refer to the forecast limit as the time when MSSS becomes zero.

By substituting Eq. (2) and (3) into Eq. (4), the estimated MSSS ($\overline{MSSS}$), as a function of ensemble size, can be described as follows.

$$\overline{MSSS}(\tau, e) = 1 - \frac{\overline{MSE_f}(\tau, e)}{MSE_c(\tau)} \approx 1 - \frac{e+1}{2e}\frac{\widetilde{MSE_f}(\tau)}{MSE_c(\tau)}$$

$$\overline{MSSS}(\tau, e) \approx 1 - \frac{e+1}{2e}\left(1 - \widetilde{MSSS}(\tau)\right) \tag{5}$$

$$\overline{MSSS}(\tau, \infty) \approx \frac{1}{2}\left(1 + \widetilde{MSSS}(\tau)\right) \tag{6}$$

Here, $\widetilde{MSSS}(\tau)$ indicates the average MSSS of individual ensemble members. From the above equations, it is shown that the skill of using e ensemble members is a function of the skill of a single-member forecast and the number of ensemble members (Eq. (5)). The maximum skill is obtained with an infinite ensemble (Eq. (6)). Figure 1a shows how the skill increases with the ensemble size for the average skill of individual members. There is a limit to the skill improvement depending on the skill of individual members, as shown in Eq. (6). Moreover, if the skill of the individual ensemble member is lower than −1, then the skill is below 0, which is the threshold of the forecast limit. In other words, to improve the prediction skill of the forecast system through the ensemble method, the predictability of individual members must be sufficiently secured.

The skill improvement ratio (SIR), which is defined by the increase of MSSS from the average skill of individual members against the theoretical maximum, is written as follows.

$$SIR(\tau, e) = \frac{MSSS(\tau, e) - \widetilde{MSSS}(\tau)}{\overline{MSSS}(\tau, \infty) - \widetilde{MSSS}(\tau)} \left(= \frac{e-1}{e} \right) \quad (7)$$

Under the perfect model condition, SIR is independent of the average skill of individual members (Figure 1b). Ten or twenty ensemble members can lead to a SIR equivalent of 90% or 95% of the theoretical maximum. This result indicates that the prediction skill of a model will be higher if more ensemble members are used. However, the result suggesting that ten or twenty ensemble members are enough to obtain the theoretical limit is consistent with the result of previous studies that discussed for a probabilistic forecast.

3. Application to the S2S Prediction System

3.1. ECMWF S2S Real-time Forecast

Among the various models participating in the S2S project, the real-time forecast data of the ECMWF prediction model were used for analysis. The model has a sufficient number of perturbed ensemble members of 51, including the control forecast. The ECMWF model data, which are provided through the S2S database, are initialized twice a week and integrated up to the forecast step of 46 days (Vitart et al., 2017).

All forecasts initialized in the period from June 2015 to May 2018 are analyzed, and those initialized in DJF (December-January-February) and JJA (June-July-August) are primarily presented. The total number of initializations is 294 forecasts, and the numbers corresponding to the DJF and JJA period is 77 and 80, respectively. Verifications are conducted for geopotential height forecasts at 50-hPa and 500-hPa representing the stratosphere and the troposphere, respectively. The region is separated into the Northern Hemisphere (30° N-90° N), the tropics (30° S-30° N), and the Southern Hemisphere (30° S-90° S).

The ECMWF model uses the Variable Resolution Ensemble System (VAREPS, Buizza et al., 2007; Vitart et al., 2008), which changes the resolution of the model from 16 km to 31 km (from 64 km to 32 km in the earlier version) at the forecast step of 15 days (10 days in the earlier version). Although the purpose of VAREPS is to save computational resources, it most likely affects deterministic predictability. Other detailed configurations of the ECMWF model are summarized in Table 1. For verification, the geopotential from ECMWF Reanalysis data-Interim (ERA-Interim, Dee et al., 2013) is interpolated to a horizontal resolution of $1.5° \times 1.5°$, and instantaneous data on 00 UTC are used. Daily climatology as the reference forecast in Eq. (4) is calculated from the 1981 to 2010 period with ERA-Interim.

3.2. Prediction Skills

Figure 2 shows the MSSS of the ECMWF model and the theoretically estimated skill over the forecast steps. Solid gray lines represent the skill of individual ensemble members, and a solid black line indicates their average. Solid colored lines represent the ensemble results of the model with a specific ensemble size. Dashed lines represent the estimated value in Eq. (5). The color of the dashed lines indicates the ensemble size; a thicker color implies that more ensemble members are used. The black line indicates the theoretical maximum.

Overall for the regions and levels, the prediction skills start from

1 at the initialization and are maintained for a few days. The MSSSs sharply decrease with the forecast steps, which depend on the choice of region and level. The general stratospheric skills continuously decrease during integration (Figure 2a-c). In the troposphere, the declining skill tends to be saturated around the forecast step of 15-20 days (Figure 2d-f).

As shown in Eq. (7), the prediction skill is expected to improve by 50% of the theoretical maximum when the ensemble size is two. For 10, 20, and 51 ensemble members, the SIR is expected to increase by 90%, 95%, and more than 98%, respectively. There are vast discrepancies between the practical skill of the model and the estimated skill in the stratosphere (Figure 2a-c). The skill is robust in the tropics, and despite the expansion of the ensemble size, the skills hardly approach the estimate, and the SIR only reaches 20-30% (Figure 3b). Ensemble size two shows a lower skill than both the estimated value and the average skill of individual ensembles. For the extratropical stratosphere in the Northern/Southern Hemisphere, the MSSS is substantially lower than the estimate during the forecast period. For the tropospheric skill evaluated with Z500, the skill of the practical model is in good agreement with the estimate in both MSSS (Figure 2d-f) and SIR (Figure 3d-f). This result indicates that approximately 20 ensemble members are needed to gain the 95% prediction skill of the maximum prediction skill in the troposphere.

These characteristics of prediction skill are also found in JJA (Figure 4). It reveals that overestimation in the stratosphere is apparent not only in the tropics but also in the summer hemisphere. These results suggest that the ensemble size effect is not consistent with level and region.

Note that there is a discontinuous change in the prediction skill at forecast step 15 days in the tropical stratosphere (Figure 3d and 4d) and the Northern Hemisphere in JJA (Figure 4a), which is due to changes in the horizontal resolution at a certain step of integration, the VAREPS. However, it is hard to recognize this discontinuity in the troposphere or other regions of the stratosphere. This strategy is useful for saving computing power without a critical decrease in prediction skill except in the tropical stratosphere.

3.3. Role of Mean Bias in Overestimation of Skill

The discrepancy between the estimated and practical skills implies that expanding ensemble size does not ensure as much of an improvement in a prediction skill as expected. As a contrapositive, this discrepancy indicates that the assumption, referred to as the perfect model condition, does not hold for the practical model, especially in the stratosphere. Since a perfect model has a fully reliable ensemble spread to observed variance, the PDFs of both the observation and the forecast should have the same statistical characteristics, such as the mean and variance. Assuming that the PDFs of the observation and forecast exhibit a normal distribution at each grid point, the entire PDF of each population can be described with only mean and variance.

The mean bias of the model is defined by the difference in means as follows.

$$MB_j(\tau) = \frac{1}{N_f}\sum_{i=1}^{N_f}\left(\tilde{f}_{ij}(\tau) - o_{ij}(\tau)\right) \quad (8)$$

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

The variance in forecast states refers to the ensemble spread

$$ENS_{s\,j}(\tau) = \frac{1}{N_f N_e}\sum_{i=1}^{N_f}\left(\sum_{e=1}^{N_e}\left(f_{ij}(\tau, e) - \tilde{f}_{ij}(\tau)\right)^2\right). \quad (9)$$

The variance in observed states can be approximated by MSE_c because a climatology can be considered as the population mean of the observations. The MSE_c at each grid point is defined as follows.

$$MSE_{c\,j}(\tau) = \frac{1}{N_f}\sum_{i=1}^{N_f}\left(c_{ij}(\tau) - o_{ij}(\tau)\right)^2 \quad (10)$$

Note that for a perfect prediction, the mean bias would be zero, and the ensemble spread would be the same as the variance in the observation (MSE_c). Figure 5 shows the distributions of the statistical characteristics of Z50 and Z500 in DJF at the forecast step of 30 days. In the stratosphere, there is a negative bias over the global area, and it is notably vast in the Northern (winter) Hemisphere. The ensemble spread and observed variance are also robust in this region. Only little bias appears in the tropical stratosphere and the summer hemisphere, where the discrepancies between practical skill and estimated skill are apparent.

In the troposphere, there is a positive bias in the summer hemisphere, but the bias is relatively small compared to that in the stratosphere, and a small ensemble spread and observed variance appear in the tropics. The mean bias and ensemble spread of the ECMWF model increases with forecast step, and the ensemble spread is similar to the observed variance in both levels when the forecast step exceeds approximately 20 days (not shown).

These results suggest that the extent of the bias does not explain the efficiency of the ensemble size effect directly. Instead, both the means of and the variances in the two PDFs should be considered simultaneously. If the difference between the means (bias) is large, and the variances in both PDFs are large enough, it is possible for the PDF of the model to capture the observed state. In contrast, if the extents of the variances are small, the forecast PDF easily deviates from the observed state even with a similar bias (Figure 6). Assuming that the ensemble spread and observed variance are the same, which seems acceptable in Figure 5, the agreement of the two PDFs is simplified to a t-test problem that rejects a mean bias of 0 as follows (s_f and s_o are the standard deviations of the forecast and observed states, respectively).

$$\left|t_j(\tau)\right| = \frac{\left|\frac{1}{N_f}\sum_{i=1}^{N_f}\left(\tilde{f}_{ij}(\tau) - o_{ij}(\tau)\right)\right|}{\sqrt{\frac{(N_f \times N_e - 1)s_f^2 + \left(N_f - 1\right)s_o^2}{N_f \times N_e + N_f - 2}}} \approx \frac{\left|MB_j(\tau)\right|}{\sqrt{\frac{\left(N_f \times N_e - 1\right)\ ENS_{s\,j}(\tau) + \left(N_f - 1\right)\ MSE_{c\,j}(\tau)}{N_f \times N_e + N_f - 2}}} \qquad (11)$$

$$\left(\nu = N_f \times N_e + N_f - 2\right)$$

Here, ν represents the degree of freedom of statics t t_j. This t-value indicates whether the mean bias is statistically significant or negligible at each grid point j (Figure 7). The PDF of both the forecast and the observed states are distinguished at the 90% confidence level ($|t| = 1.65$) in the tropical stratosphere at the forecast step of 30 days. In mid-high latitudes, the t-value is smaller than 1.28, which rejects hypothesis at the 80% confidence level. This t-value increases with the forecast step (not shown). At the forecast step of 15 days, the region of rejecting the hypothesis at the 90% confidence level is narrower than 30 days. At the forecast step of 46 days, which is the maximum forecast step for the ECMWF real-time forecast data, the t-value increases over the 95% confidence level ($|t| = 1.96$). In the troposphere, the extent of the t-value is small enough that the mean bias can be considered as 0. This result suggests that expanding ensemble size may not necessarily lead to an improvement in the predictability, even if the absolute extent of the bias is vast, depending on whether the PDFs coincide with each other. Furthermore, a statistically significant bias cannot be solved through only an ensemble system, which means that the physical and dynamical processes in the model should be developed for the area where bias occurs.

One of the simplest ways to prevent the problem of a statistically significant mean bias is to subtract the model bias from the forecast value, referred to as bias correction. Figures 8 and 9 show the MSSS in DJF and JJA, respectively, against forecast step for f', which is the model result after bias correction as follows.

$$f'_{ij}(\tau, e) = f_{ij}(\tau, e) - MB_j(\tau)$$

Compared with the results in Figure 2, these results show that not only an improvement in the MSSS in each area but also a practical skill consistent with theoretical skill. In particular, these changes in prediction skill and ensemble size effect are most pronounced in the S2S time-scale (forecast step > 10 days) for the stratosphere. In short, the simple bias correction process removing the model bias reduces the t-value (Eq. (11)), which indicates improvement of the effectiveness of the ensemble size effect so that improving the skill to the same level as the theory. Furthermore, this result implies that a bias correction with weak physical or dynamical meaning can correct the PDF of the forecast state from a statistical perspective.

3.4. Vertical and Seasonal Extension

To confirm whether the above discussion can be applied to only some levels and seasons, the same t-value analysis is extended by level and season. First, to ensure that 50-hPa and 500-hPa geopotential height forecasts are representative of the stratospheric and tropospheric skills, respectively, the mean t-value is obtained by averaging each region at all altitudes[①] provided by the S2S database (Vitart et al., 2017). Figure 10 shows that as the forecast step increases, a statistically significant mean bias appears in the tropical and summer hemispheric stratosphere over 100-hPa pressure level. This mean bias is not only related to the sudden skill drop in the summer stratosphere reported in previous studies (Domeisen et al., 2019; Son et al., 2019) but also shows that the substantially low ensemble prediction skill of the 50-hPa geopotential height forecasts

---

① 1000/925/850/700/500/300/200/100/50/10-hPa pressure levels is consistent in the upper stratosphere.

However, unlike that at 50-hPa, the t-value at the 10-hPa tends to show a maximum at 10-15 forecast days and gradually decreases in the summer hemisphere (Figure 10c). Although the extent of the mean bias at this forecast step is smaller than that at a longer lead time, the ensemble spread is relatively smaller than the observed variance (not shown). This result indicates that the ensemble spread grows too slowly to represent the model uncertainty. The decline in the efficiency of the ensemble size effect in the upper stratosphere is affected by the mean bias, and the ensemble spread also plays an important role in the summer hemispheric 10-hPa level. The detailed role of ensemble spread needs to be explored in a future study.

Figure 11 shows the t-value averaged over three months in each area. The stratospheric t-value is substantially higher than the t-value of the troposphere that is less than $|t| = 0.67$, which indicates that the model mean bias is significant at the 50% confidence level, which is consistent with the results of previous chapters. In the stratosphere, the Northern Hemispheric t-value is higher than the Southern Hemispheric t-value (Figure 11a and c). The tropospheric t-values are substantially close to zero, which implies that the bias in the troposphere is considered negligible.

In regard to the seasonal variation, both extratropical stratospheric domains represent high t-values except in the winter season. As the forecast step increases, the seasonal variation in the t-value becomes apparent, which can be seen as the influence of the mean bias that increases with the forecast step. Note that seasonal variation in the extratropical stratospheric t-value is more affected by the observed variance and corresponding ensemble spread than the extent of the mean bias itself, as shown in Figure 6 (Figure 11a and c). On the other hand, the tropical stratosphere, which has a relatively small natural variability, represents high t-values, and the seasonal variation in the mean bias directly affects the change in t-value. As a result, the t-value tends to increase during the transition seasons. These results indicate that the decrease in the ensemble size effect due to the mean bias in the tropical and summer hemispheric stratosphere also appears in other seasons, and its magnitude is dependent on the seasons either. As shown in chapter 3.3, the effect of mean bias can be partially removed through the bias correction; however, fundamentally, the cause of the mean bias in each domain should be explored and improved by model development through future research.

4. Summary and Conclusions

In this study, the deterministic prediction skill is theoretically estimated to optimize the ensemble size for efficient prediction. When assuming a perfect model condition, it is possible to estimate the prediction skill from the skill of individual members and the number of ensemble members. The SIR, which is defined by the increase in MSSS against the theoretical maximum (or infinite ensemble size), is independent of the skill of individual members. When determining the optimal number of ensemble members based on skill improvement, more than 10 or 20 ensemble members are needed to obtain skill improvement equivalent to 90% or 95% of the theoretical maximum.

This theoretical approach is applied to the ECMWF S2S real-time forecast data that consist of 51 ensemble members. In the troposphere, the prediction skill of the model is in good agreement with the theoretical estimation. It implies that approximately 20 ensemble members are needed to establish the ensemble system, which has a SIR of 95% of the theoretical maximum. On the other hand, the stratospheric prediction skill underperforms compared to the estimate, especially in the tropics and summer hemisphere.

If the model has a statistically significant bias from the reference, the perfect model condition cannot hold. The discrepancies between the estimated and actual skills are ascribed to the bias, and their variances have critical roles. These discrepancies are assured and partly solvable through the bias correction, which is the process that eliminates bias from the forecasts.

The above discussions are also extended to level and season. It turns out that a bias affecting the ensemble size effect consistently appears in the upper stratosphere (above 100-hPa pressure level). The 10-hPa t-value, which shows the maximum at 10 to 15 forecast days, suggests that the ensemble spread is also crucial in this level and that its role needs to be explored in a future study. The stratospheric bias is considered statistically significant overall seasons, though the bias can be assumed to be zero in the troposphere.

Even though expanding the ensemble size is a relatively straightforward and ensuring way to improve prediction skills, this study suggests that expanding the ensemble size is not a sufficient solution. Ultimately, elimination of bias should be performed through developing the model itself, which is physically and dynamically more meaningful than the bias correction. Moreover, the SIR (Eq. (7)) used as an indicator of the efficiency of the ensemble size effect is not related to the skill of the individual ensemble members; however, the MSSS (Eq. (5)) which represents the predictability of the prediction system is eventually determined by the average skill of individual ensembles. These results suggest that it is essential to increase the performance of the model itself as well as to expand the ensemble size to improve the predictability of the system.

References

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch (1990) Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Q. J. R. Meteor. Soc.,* **116,** 867-912, DOI: 10.1002/qj.49711649405

Buizza, R., J.-R. Bidlot, N. Wedi, M. Feuntes, M. Hamrud, G. Holt, and, F. Vitart (2007) The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Q. J. R. Meteor. Soc.,* **133,** 681-695, DOI: 10.1002/qj.75

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Holm, L. Isaksen, P. Kallberg, M. Kohler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz , J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thepaut and F. Vitart (2013) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteor. Soc.,* **137,** 553-597, DOI: 10.1002/qj.828

Deque, M. (1997) Ensemble size for numerical seasonal forecasts. *Tellus,* **49A,** 74-86, DOI: 10.1034/j.1600−0870.1997.00005.x

Domeisen, D. I. V., A. H. Butler, A. J. Charlton-Perez, B. Ayarzaguena, M. P. Baldwin, E. Dunn-Sigouin, J. C. Furtado, C. I. Garfinkel, P.

Hitchcock, A. Y. Karpechko, H. Kim, J. Knight, A. L. Lang, E. -P. Lim, A. Marshall, G. Rolff, C. Schwartz, I. R. Simpson, S. -W. Son, and M. Taguchi (2019) The role of the stratosphere in sub-seasonal to seasonal prediction. Part I: Predictability of the stratosphere. *submitted.*

Gneiting T. and A. E. Raftery (2005) Weather forecasting with ensemble methods. *Science,* **310,** 248-249, DOI: 10.1126/science.1115255

Kumar, A., A. G. Barnston, and M. Hoerling (2001) Seasonal predictions, probabilistic verification and ensemble size. *J. Clim.,* 14, 1671-1676, DOI: 10.1175/1520−0442(2001)014<1671:SPPVAE>2.0.CO;2

Leith, C. E. (1974) Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102.6,** 409-418, DOI: 10.1175/1520−0493(1974)102<0409:TSOMCF>2.0.CO;2

Lim, Y., S.-W. Son, and D. Kim (2018) MJO prediction skill of the subseasonal-to-seasonal (S2S) models. *J. Clim.,* **31,** 4075-4094, DOI: 10.1175/JCLI−D−17−0545.1

Murphy, A. H. (1988a) Skill scores base on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417-2424, DOI: 10.1175/15200493(1988)116<2417:SS BOTM>2.0.CO;2

Murphy, A. H. and E. S. Epstein (1989) Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572-581, DOI: 10.1175/1520−0493(1989)117<0572:SSACCI>2.0.CO;2

Murphy, J. M. (1988b) The impact of ensemble forecasts on predictability. *Q. J. R. Meteor. Soc.,* **114,** 463-493, DOI: 10.1002/qj.49711448010

Richardson, D. S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteor. Soc.,* 127, 2473-2489, DOI: 10.1002/qj.49712757715

Robertson, A. W. and F. Vitart (Eds.) (2018) Sub-seasonal to seasonal prediction: The gap between weather and climate forecasting. Amsterdam, AE: Elsevier.

Son, S.-W., H. Kim, K. Song, S.-W. Kim, P. Martineau, Y.-K. Hyun, and Y. Kim (2019) Extratropical prediction skill of the Subseasonal-to-Seasonal (S2S) prediction models. *submitted.*

Vitart, F., R. Buizza, M. A. Balmaseda, G. Balsamo, J.-R. Bidlot, A. Bonet, M. Fuentes, A. Hofstadler, F. Molteni, and T. N. Palmer (2008) The new VarEPS-monthly forecasting system: A first step towards seamless prediction. *Q. J. R. Meteor. Soc.,* **134,** 1789-1799, DOI: 10.1002/qj.322

Vitart, F., A. W. Robertson, D. L. T. Anderson (2012) Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *Bull. World Meteor. Org.,* **61.2,** 23-28.

Vitrat, F. (2017) Madden-Julian Oscillation prediction and teleconnections in the S2S database. *Q. J. R. Meteor. Soc.,* **143,** 2210-2220, DOI:10.1002/qj.3079

Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Deque, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H.−S. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najmjm, Y. Nie, M. Rixen, A. W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D.−J. Won, H. Xiao, R. Zaripov, and L. Zhang (2017) The subseasonal to seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.,* **98,** 163-173, DOI: 10.1175/BAMS−D−16−0017.1

World Meteorological Organization (2013) Sub-seasonal to seasonal prediction research implementation plan. [Available online at https://www.wmo.int/pages/prog/arep/wwrp/new/documents/S2S_Implem_plan_en.pdf], Geneva, Switzerland

Table

Table 1. Information about the ECMWF model (on the fly) configuration.

Update Date	Model Version	Time Range	Ens. Size	Atm. Resolution	Ocean Resolution	Active Sea Ice
14/05/2015	CY41R1	d 0-46	51	32km(~10d) 64km(10d~)	1 degree	NO
08/03/2016	CY41R2	d 0-46	51	16km(~15d) 31km(15d~)	1 degree	NO
22/11/2016	CY43R1	d 0-46	51	16km(~15d) 31km(15d~)	1/4 degree	YES
11/07/2017	CY43R3	d 0-46	51	16km(~15d) 31km(15d~)	1/4 degree	YES

Figures

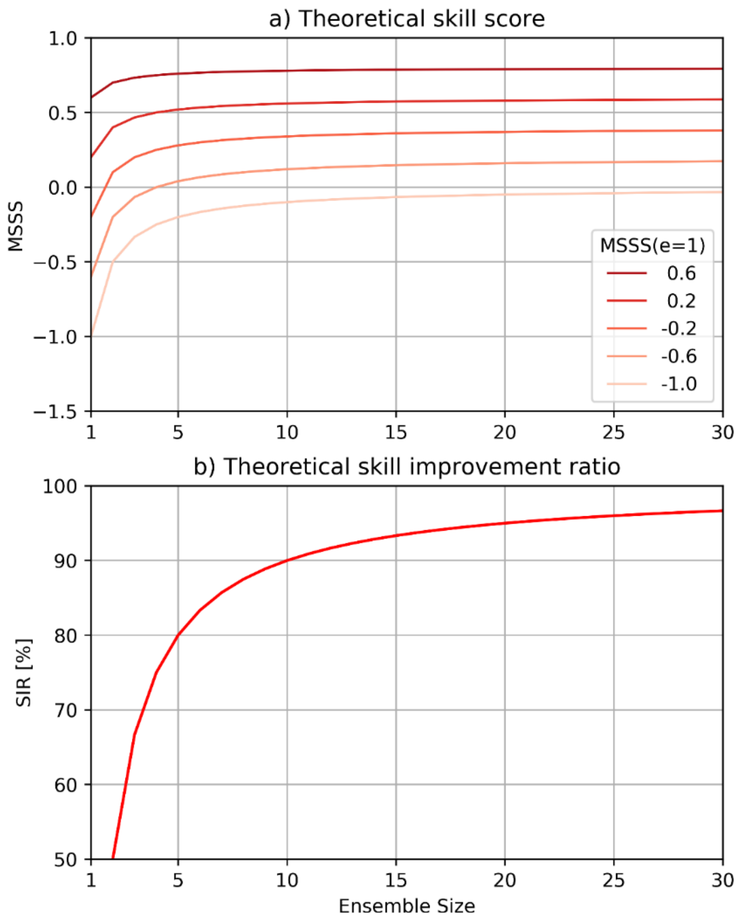

Figure 1. Theoretical MSSS (a) and SIR (b) of a perfect model with ensemble size.

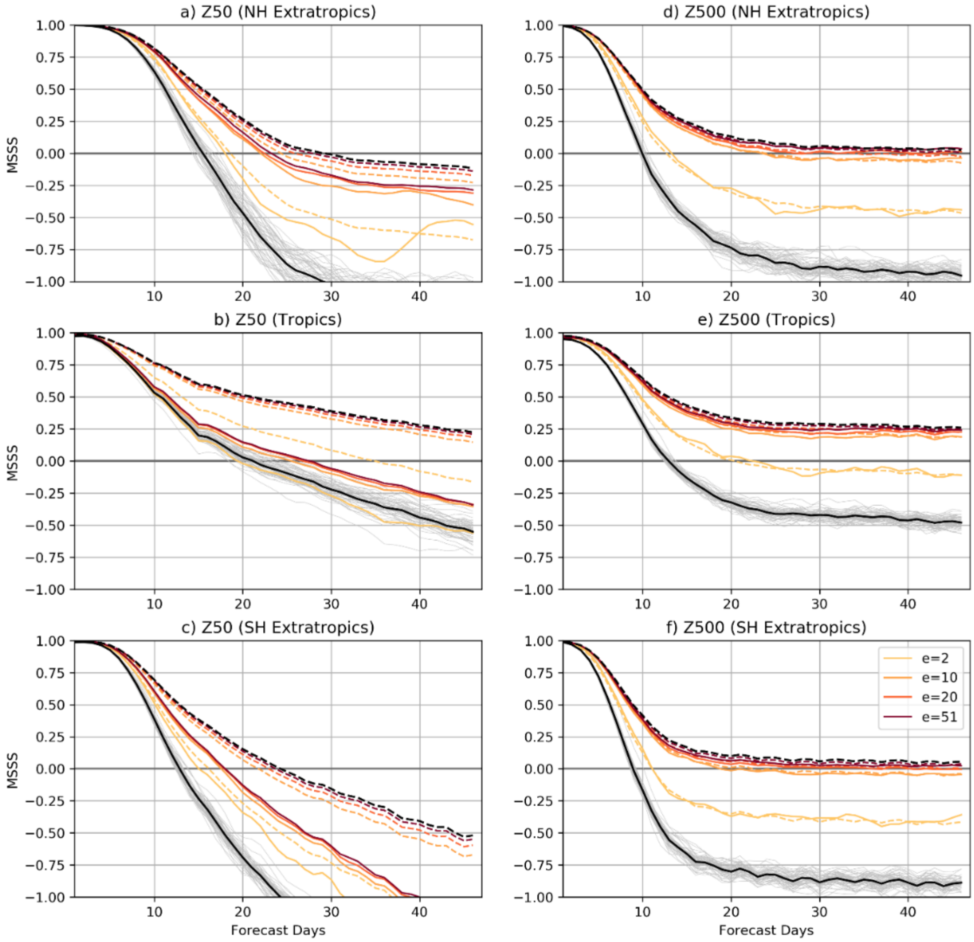

Figure 2. MSSS of individual ensemble members (gray lines) and the ensemble results (colored solid lines) and estimated MSSS with ensemble size (colored dashed lines) of the 50-hPa (left) and 500-hPa (right) geopotential height forecasts initialized in DJF. The averaged skill of individual ensemble members and the theoretical skill limit are represented by a solid black line and a dashed black line, respectively.

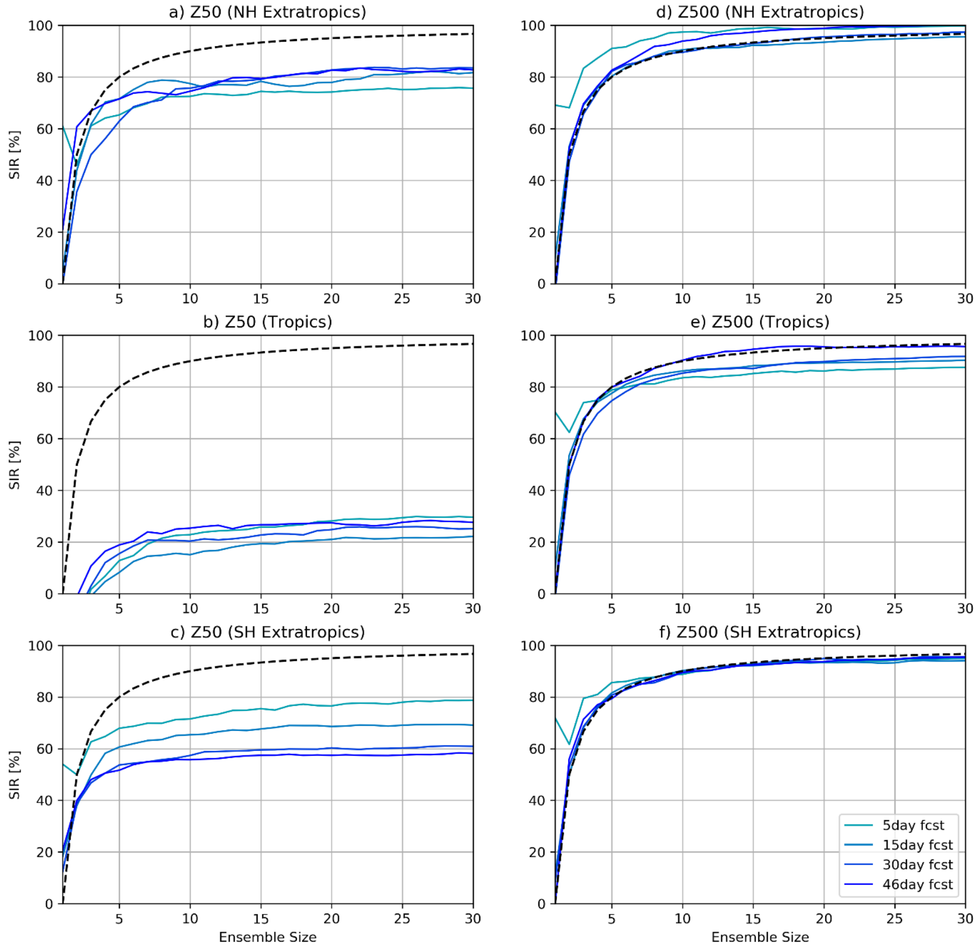

Figure 3. SIR of the 50-hPa (left) and 500-hPa (right) geopotential height forecasts initialized in DJF. The color of the lines indicates forecast steps from initialization. A black dashed line represents the SIR of a perfect model.

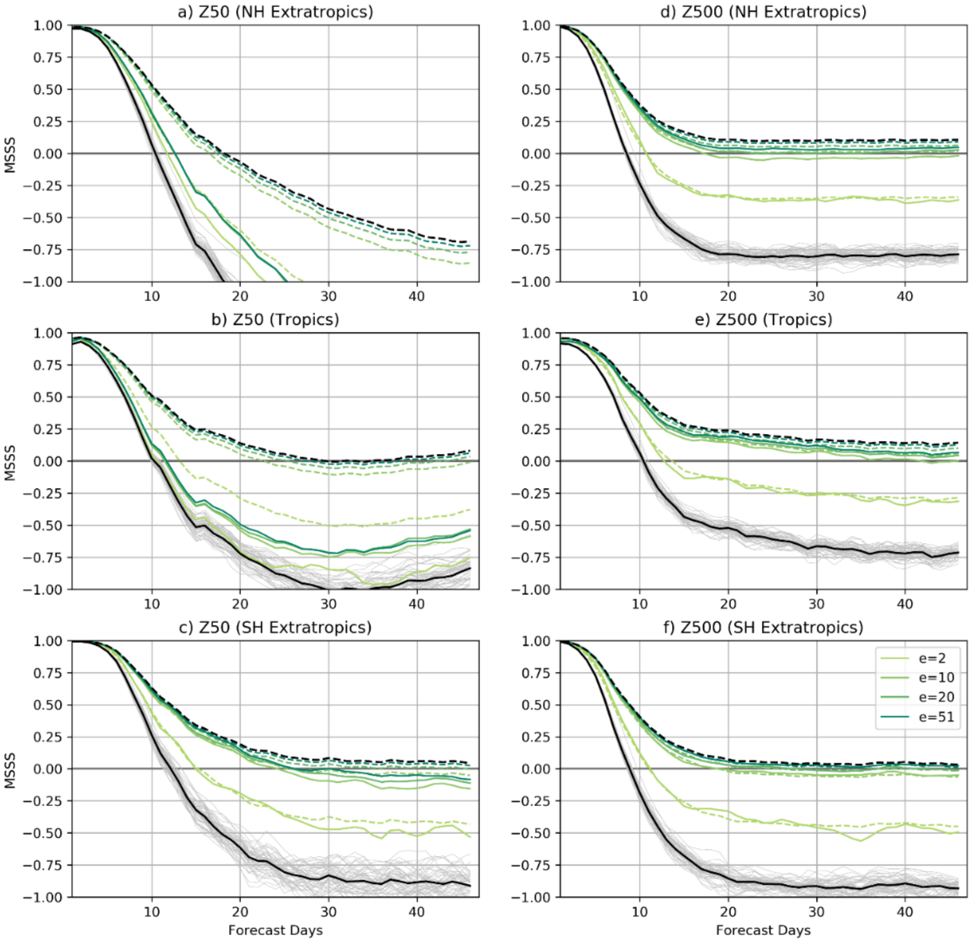

Figure 4. Same as Figure 2, but for forecasts initialized in JJA.

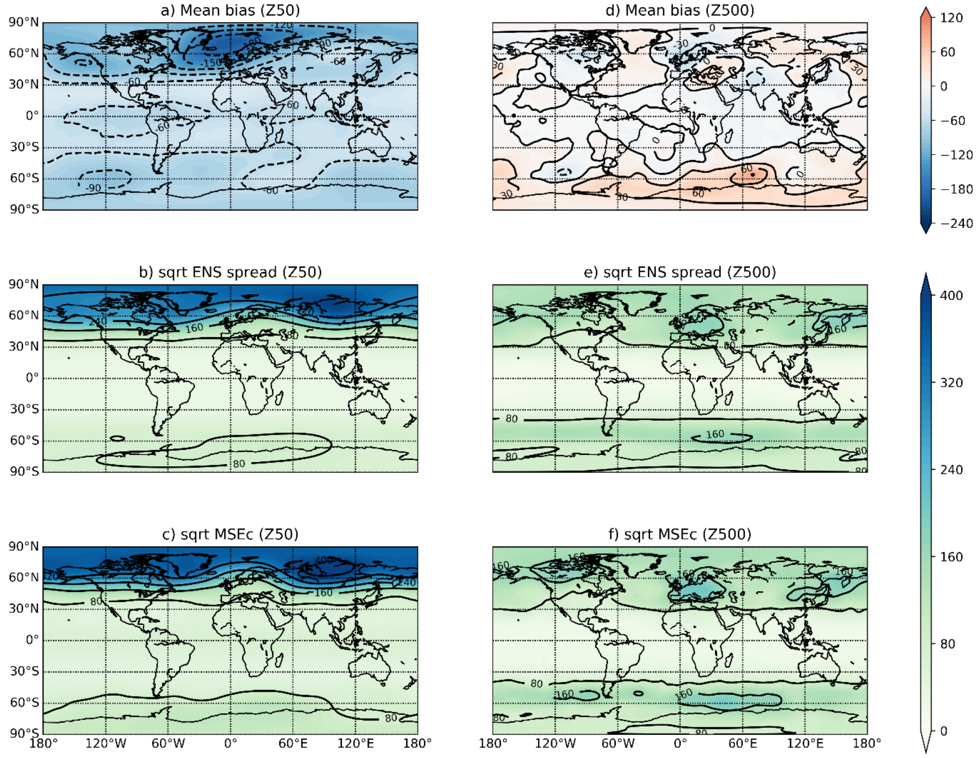

Figure 5. Mean bias (top), square root ensemble spread (middle), and square root MSE_c distribution of the 50-hPa (left) and 500-hPa (right) geopotential height forecasts at the forecast step of 30 days for those initialized in DJF.

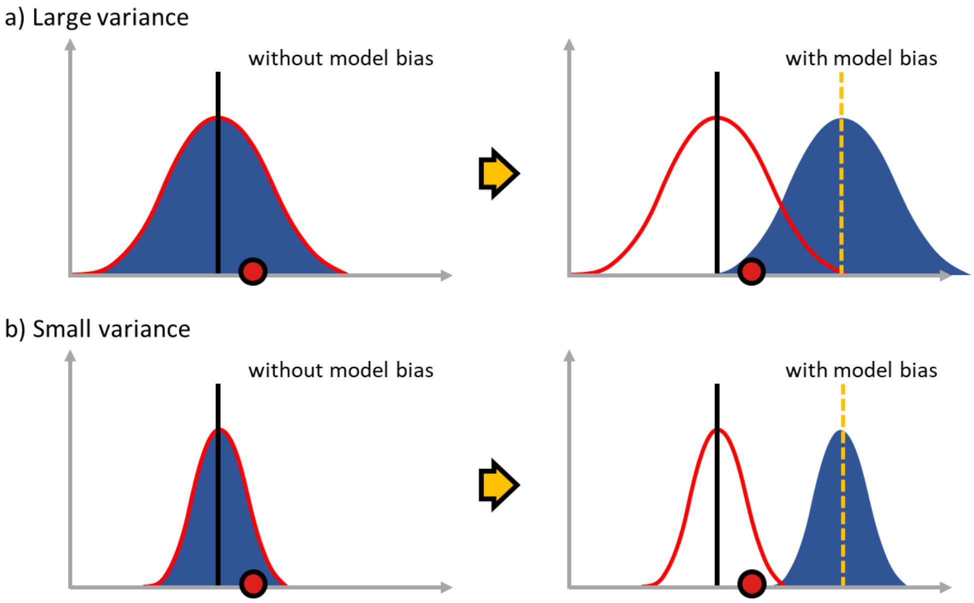

Figure 6. Schematic diagram of the probability density function (PDF) of the observed state (red line) and model ensembles (blue shaded) when the PDFs have a large (a) and small (b) variance. Without the model bias, the model ensemble PDF can capture the observation regardless of the size of the variance (left); however, when bias occurs, it is relatively hard for the PDF with the smaller variance to capture the observation (right).

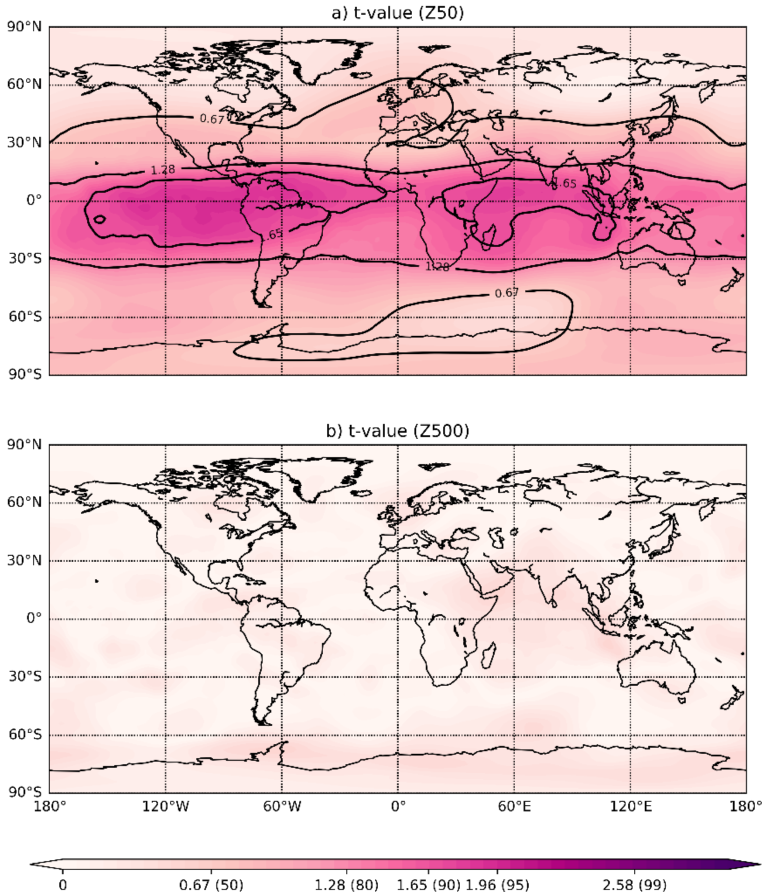

Figure 7. Distribution of t-value for the mean bias of the 50-hPa (above) and 500-hPa (below) geopotential height forecasts at the forecast step of 30 days for those initialized in DJF.

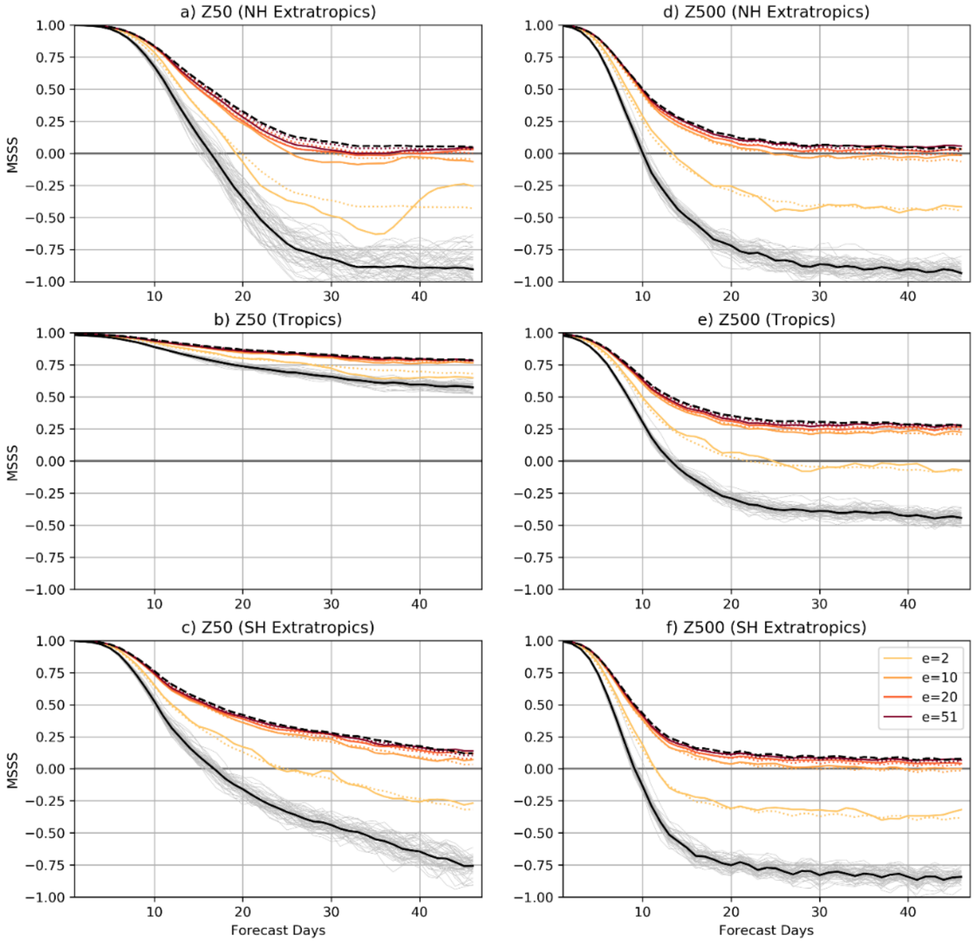

Figure 8. Same as Figure 2, but for the result with bias correction.

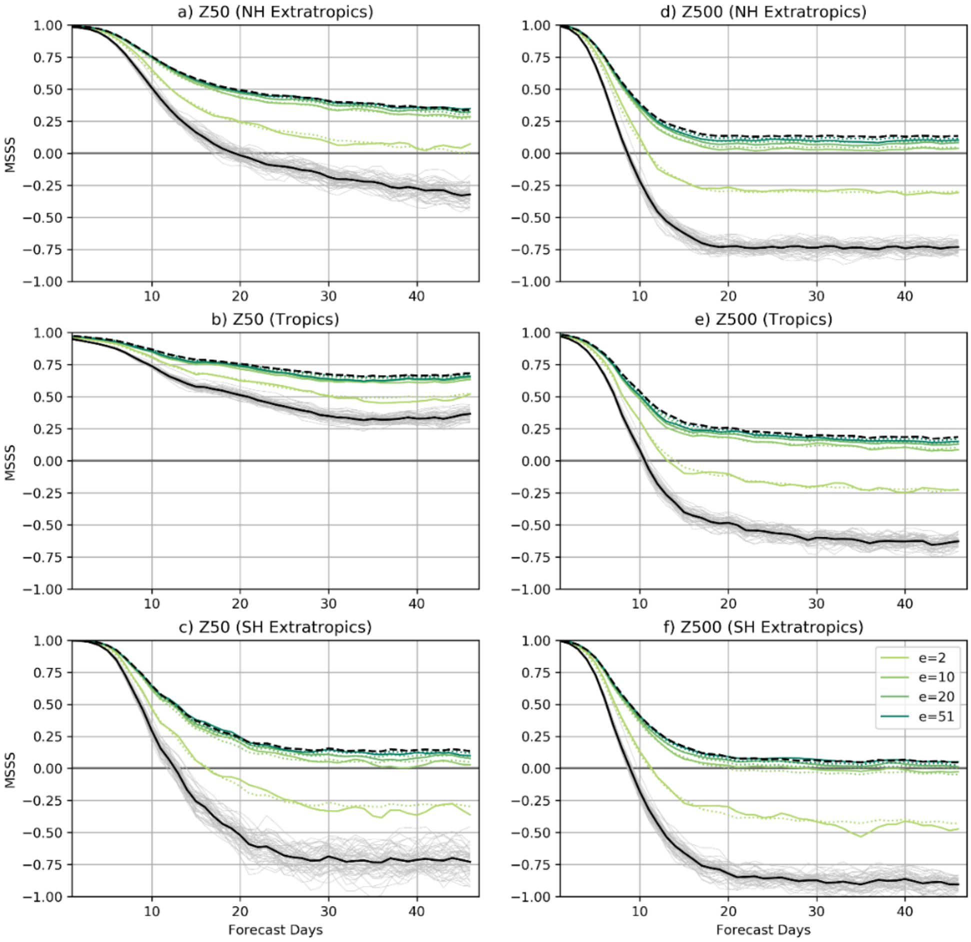

Figure 9. Same as Figure 4, but for the result with bias correction.

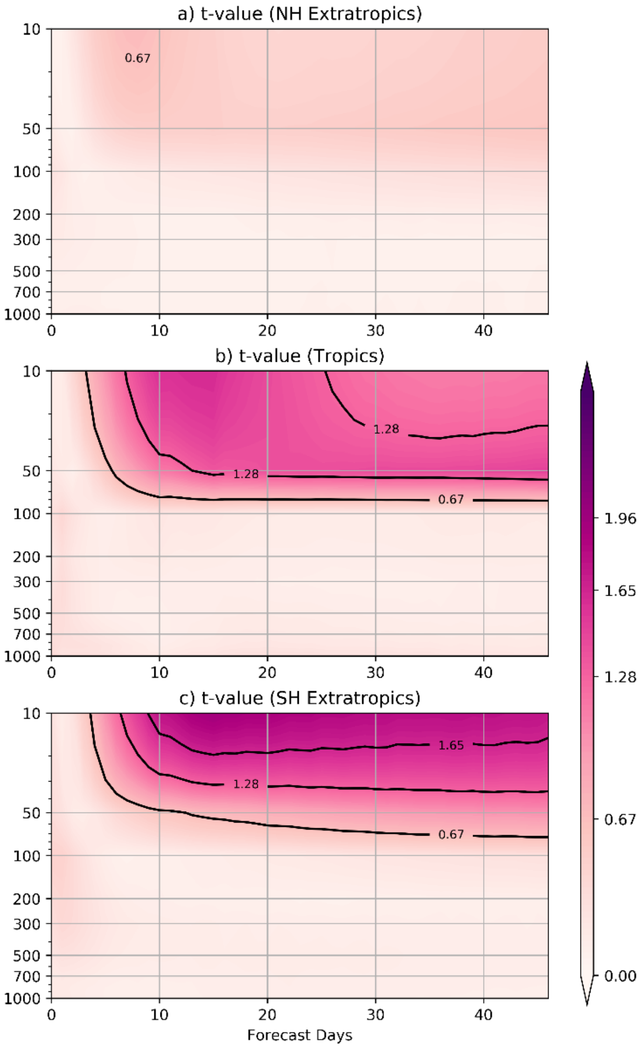

Figure 10. Vertical distribution of t-value for the mean bias of geopotential height forecasts against forecast steps for those initialized in DJF.

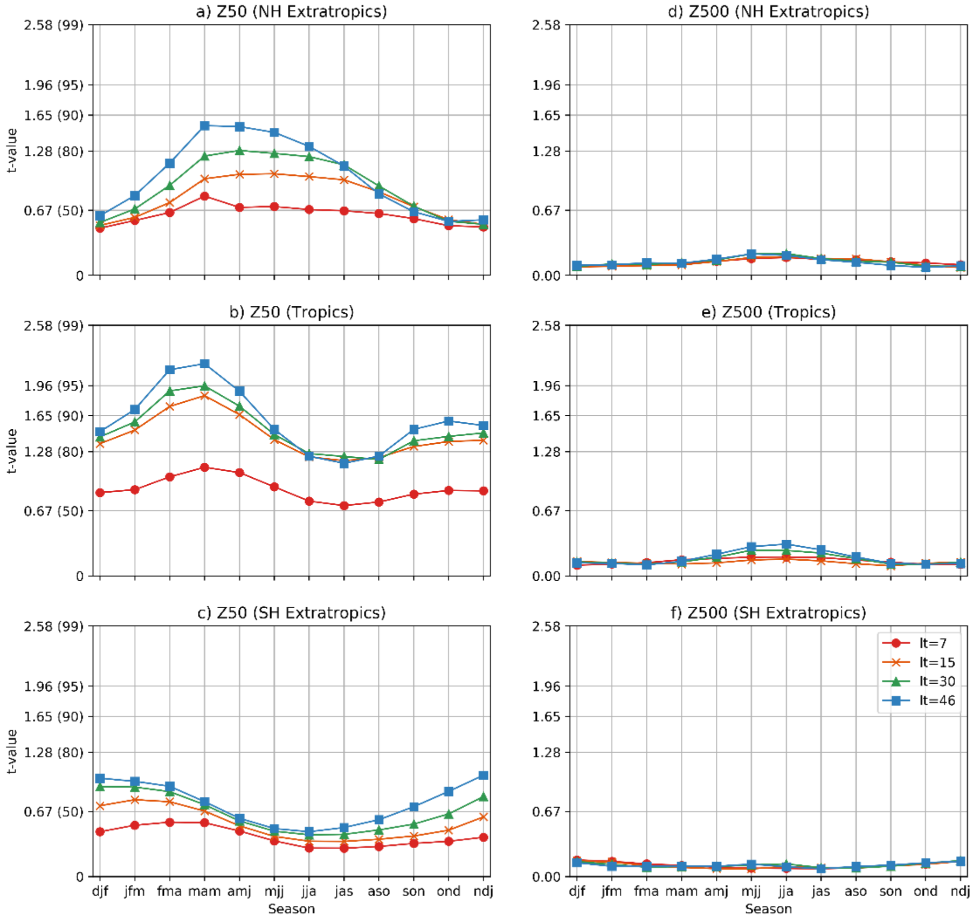

Figure 11. Monthly difference in averaged t-value for the mean bias of the 50-hPa (left) and 500-hPa (right) geopotential height forecasts.

Abstract

계절내 및 계절(S2S) 예측을 위한 적절한 앙상블 크기 설정에 대한 연구

김 상 욱
지구환경과학부
석사과정
서울대학교

본 연구는 계절내 및 계절(S2S) 예측을 위한 최적의 앙상블 크기와 그 결정에 영향을 미치는 요인에 대해 다루고 있다. 바이어스가 없고 관측의 분산에 잘 부합하는 앙상블 스프레드를 갖춘 완벽한 모델에서, 평균제곱기술(Mean Square Skill Score; MSSS)로 정량화되는 모델의 예측성은 앙상블 크기가 증가함에 따라 항상 향상됩니다. 이러한 이상적인 예측시스템에서 10-20개의 앙상블 멤버가 있으면 이론적으로 얻을 수 있는 최대(무한 개 앙상블 멤버)의 90에서 95%에 해당하는 예측성 향상을 기대할 수 있다. 51개의 앙상블 멤버를 가진 유럽중기예측센터(European Centre for Medium-Range Weather Forecasts; ECMWF)의 S2S 실시간 예측 자료를 위의 이론적 추정과 비교하였다. 대류권의 예측성을 나타내는 500-hPa 지위고도의 MSSS는 이론적인 추정치와 잘

일치하며, 이는 10-20개의 멤버로 이론적인 최대치에 근접한 예측성 향상을 얻을 수 있음을 의미한다. 그러나 50-hPa 지위고도의 MSSS로 평가된 성층권 예측성은 이론적 추정에 비해 상당히 낮은 것으로 나타났다. 특히 열대 성층권 영역에서는 모든 앙상블 멤버를 사용했을 때에도 예측성 향상이 이론값의 20-30% 수준에 불과했다. 성층권에서 나타나는 과대추정 경향은 주로 모델의 평균 바이어스에 기인한다. 이 같은 앙상블 크기 효과에 대한 바이어스의 영향은 앙상블 스프레드와 관측의 분산에 크게 영향을 받는다. 단순히 바이어스를 제거하는 바이어스 보정(bias correction)을 통해 성층권에서의 과대추정을 완화할 수 있으며, 대류권과 마찬가지로 10-20개를 최적 앙상블 규모로 취할 수 있다. 전체 고도와 계절에 대한 확장을 통해, 열대 및 여름반구 상부 성층권에서 동일하게 바이어스의 영향이 극명하게 나타남을 확인할 수 있다.

Keywords : 계절내 및 계절(S2S) 예측, 앙상블 예측 시스템(EPS),
앙상블 크기 효과, 최적 앙상블 크기, 예측성
Student Number : 2017-20309