creative commons
COMMONS DEED

저작자표시-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사학위논문

확률론적 홍수예측 정확도에 대한 상관 앙상블 시나리오의 효과 분석

Effects of Correlated Ensemble Scenarios on Accuracy of Probabilistic Flood Forecasting

2013년 12월

서울대학교 대학원

건설환경공학부

서 영 호

초 록

전 세계적으로 홍수예측 시스템 분야에서 확정적인 예측보다는 앙상블 예측 시스템(Ensemble Prediction System, EPS)에 대한 연구가 증가하고 있다(Cloke and Pappenberger, 2009). 또한 수문학 분야에서 이를 이용하여 유량예측을 하는 앙상블 유량예측(Ensemble Streamflow Prediction, ESP)에 대한 연구가 활발히 진행되어 왔다. 발생 확률의 범위를 표현하기 위해서 확률예측이 가지는 전체 불확실성을 분석할 필요가 있으며, 이를 위하여 많은 불확실성 분석 연구가 진행되어왔다.

많은 앙상블 예측 시스템들은 동등한 확률을 가지는 앙상블 멤버로 구성되도록 설계되어 있다(Cloke and Pappenberger, 2009). 따라서 발생 확률의 범위를 표현하기 위하여 적절한 수의 앙상블 시나리오가 설정되어야 한다. 하지만 확률 홍수예측을 위하여 적정한 수의 시나리오를 결정하는 것은 어려운 일이며, 여러 연구가 진행되었으나 적절한 앙상블 시나리오 수에 대한 합의점을 보여주지 못하고 있다.

또한 앙상블 시나리오들이 동등한 확률을 가진다는 것은 시나리오 간 독립성을 가정하고 있음을 의미한다. 앙상블 시나리오 간 상관성이 존재하여 불확실성이 증가되고 신뢰성이 감소될 수 있다. 하지만 앙상블 시나리오 간 상관성이 미치는 영향에 대한 연구는 검토가 되지 못하고 있다. 따라서 본 연구에서는 시나리오 간 상관성 정도와 시나리오 수가 예측에 미치는 영향을 분석하였다. 여기서 분석의 일반화를 위하여 각 시나리오가 표준정규분포를 따른다고 가정하였다. 앙상블 간 상관계수는 0부터 0.9까지 6가지, 그리고 앙상블 시나리오 개수는 3부터 100까지 10가지로 설정하였다. 이를 이용하여 다변량 상관 난수를 생성하여 상관 앙상블(correlated ensemble)로 사용하였다. 그리고 정확도를 조정하여 관측값을 생성

한 뒤 홍수예측의 정확도별로 구분하여 시나리오 간 상관성과 개수가 미치는 효과를 분석하였다.

예측 정확도 분석을 위해 확률적 예측 평가방법인 브라이어 스코어(Brier Score, BS)와 확정적 예측 평가방법인 D^2(squared dispersion)와 SE(Squared Error) 사용하여 결과를 분석하였다. 정확도가 높지 못한 경우에 상관계수가 작을수록 BS값이 작은 것을 확인할 수 있다. 반대로 정확도가 높은 경우에는 상관계수가 작을 경우에 BS값이 큰 것을 확인하였다. 따라서 정확한 유량예측을 하기 어려운 경우에는 시나리오 간 독립성이 보장되고 시나리오 수가 많을수록 앙상블 예측이 효과적임을 알 수 있었다. 이와 반대로 정확한 유량예측을 수행할 수 있는 경우에는 전자에 비하여 앙상블 예측의 효율성이 저하되었다. 또한 앙상블 시나리오 수의 증가가 예측의 정확도를 낮출 가능성이 있음을 확인하였다.

BS를 사용하여 앙상블 예측 시스템이 효과적으로 향상되는 앙상블 멤버 수를 찾으려고 하였다. 유효 앙상블 시나리오 수는 적절한 예측의 정확도를 가지는 최소한의 시나리오 수로 정의하였다. 유효 앙상블 시나리오 수를 제시하기 위하여 2가지 방법을 사용하였다. 첫 번째 방법은 BS의 최대값으로부터 BS가 90% 향상된 지점의 시나리오 수를 추정하였다. 그리고 두 번째 방법은 각 구간의 BS 기울기를 구하여 최대 기울기의 5% 구간을 탐색하였다. 그 결과 앙상블이 시나리오 수 증가와 함께 정확도가 높아지는 경우 20~25개가 적합하다는 결과를 얻을 수 있었다.

주요어: 앙상블 유량예측, 상관성, 다변량 상관 난수 생성, 유효 앙상블 시나리오 수, Brier score

학 번: 2012-20904

목 차

초 록 ····· i
목 차 ····· iii
표 목차 ····· v
그림 목차 ····· vi
기호 ····· x

제 1 장 서 론 ····· 1
1.1 연구의 배경 및 필요성 ····· 1
1.2 연구의 목적 ····· 4
1.3 논문의 구성 ····· 4

제 2 장 연구동향 ····· 5
2.1 시·공간 상관관계와 수문 분야 ····· 5
2.2 앙상블 유량예측 ····· 7
2.3 앙상블 불확실성 분석 ····· 10

제 3 장 연구 방법 ····· 12
3.1 앙상블 생성 ····· 12
3.1.1 다변량 상관관계(Multivariate Correlation) ····· 12
3.1.2 상관 난수 생성(Correlated Random Number Generation) ····· 20
3.1.3 상관 앙상블 생성 및 관측시나리오 생성 ····· 27
3.2 앙상블 평가방법 ····· 33
3.2.1 확정적 예측 평가방법 ····· 33
3.2.2 다중범주 확률예측 평가방법 ····· 36
3.2.3 유효 앙상블 시나리오 수 추정방법 ····· 38

제 4 장 적용 및 결과 ····· 39

4.1 개요 ····· 39
4.2 변동계수 효과 분석 ····· 41
4.3 앙상블 시나리오 간 상관성 효과분석 ····· 47
4.3.1 확정적 예측 평가 ····· 47
4.3.2 다중범주 확률예측 평가 ····· 83
4.3.3 앙상블 시나리오 간 상관성 효과분석 결과 ····· 111
4.4 유효 앙상블 시나리오 수 추정 ····· 118
4.4.1 90% 향상 구간 추정 ····· 118
4.4.2 구간별 향상률 비교 ····· 122

제 5 장 결론 및 향후 연구 ····· 124
5.1 요약 및 결론 ····· 124
5.2 향후 연구 ····· 128

참고문헌 ····· 130

Abstract of the thesis ····· 139

표 목차

Table 4.1 The effective number of ensemble scenarios ····· 121
Table 4.2 The slope of Brier score curve between each interval ····· 123

그림 목차

Fig 3.1 Vectors and in 3D-space ····· 15
Fig. 3.2 Coefficient of variation of annual flows vs mean annual runoff ····· 29
Fig. 3.3 Procedure of study ····· 32
Fig. 4.1 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.1$ ····· 42
Fig. 4.2 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.3$ ····· 43
Fig. 4.3 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.5$ ····· 44
Fig. 4.4 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.7$ ····· 45
Fig. 4.5 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.9$ ····· 46
Fig. 4.6 Squared error of ensemble mean forecasts ($\rho_e = 0$) ····· 49
Fig. 4.7 Squared error of ensemble mean forecasts ($\rho_e = 0.1$) ····· 49
Fig. 4.8 Squared error of ensemble mean forecasts ($\rho_e = 0.3$) ····· 50
Fig. 4.9 Squared error of ensemble mean forecasts ($\rho_e = 0.5$) ····· 50
Fig. 4.10 Squared error of ensemble mean forecasts ($\rho_e = 0.7$) ····· 51
Fig. 4.11 Squared error of ensemble mean forecasts ($\rho_e = 0.9$) ····· 51
Fig. 4.12 Mean squared dispersion (summary) ····· 53
Fig. 4.13 Mean squared dispersion(ensemble members, 3) ····· 53
Fig. 4.14 Mean squared dispersion(ensemble members, 5) ····· 54
Fig. 4.15 Mean squared dispersion(ensemble members, 7) ····· 54
Fig. 4.16 Mean squared dispersion(ensemble members, 9) ····· 55
Fig. 4.17 Mean squared dispersion(ensemble members, 12) ····· 55
Fig. 4.18 Mean squared dispersion(ensemble members, 15) ····· 56
Fig. 4.19 Mean squared dispersion(ensemble members, 20) ····· 56
Fig. 4.20 Mean squared dispersion(ensemble members, 30) ····· 57
Fig. 4.21 Mean squared dispersion(ensemble members, 50) ····· 57
Fig. 4.22 Mean squared dispersion(ensemble members, 100) ····· 58
Fig. 4.23 Mean squared dispersion(uncorrelated, 0) ····· 58
Fig. 4.24 Mean squared dispersion($\rho_e = 0.1$) ····· 59
Fig. 4.25 Mean squared dispersion($\rho_e = 0.3$) ····· 59

Fig. 4.26 Mean squared dispersion($\rho_e = 0.5$) ····· 60
Fig. 4.27 Mean squared dispersion($\rho_e = 0.7$) ····· 60
Fig. 4.28 Mean squared dispersion($\rho_e = 0.9$) ····· 61
Fig. 4.29 SE of the ensemble mean forecasts ($\rho_e = 0,\ \rho_a = 0.1$) ····· 64
Fig. 4.30 SE of the ensemble mean forecasts ($\rho_e = 0.1,\ \rho_a = 0.1$) ····· 64
Fig. 4.31 SE of the ensemble mean forecasts ($\rho_e = 0.3,\ \rho_a = 0.1$) ····· 65
Fig. 4.32 SE of the ensemble mean forecasts ($\rho_e = 0.5,\ \rho_a = 0.1$) ····· 65
Fig. 4.33 SE of the ensemble mean forecasts ($\rho_e = 0.7,\ \rho_a = 0.1$) ····· 66
Fig. 4.34 SE of the ensemble mean forecasts ($\rho_e = 0.9,\ \rho_a = 0.1$) ····· 66
Fig. 4.35 SE of the ensemble mean forecasts ($\rho_e = 0,\ \rho_a = 0.3$) ····· 68
Fig. 4.37 SE of the ensemble mean forecasts ($\rho_e = 0.1,\ \rho_a = 0.3$) ····· 68
Fig. 4.38 SE of the ensemble mean forecasts ($\rho_e = 0.3,\ \rho_a = 0.3$) ····· 69
Fig. 4.39 SE of the ensemble mean forecasts ($\rho_e = 0.5,\ \rho_a = 0.3$) ····· 69
Fig. 4.40 SE of the ensemble mean forecasts ($\rho_e = 0.7,\ \rho_a = 0.3$) ····· 70
Fig. 4.41 SE of the ensemble mean forecasts ($\rho_e = 0.9,\ \rho_a = 0.3$) ····· 70
Fig. 4.42 SE of the ensemble mean forecasts ($\rho_e = 0,\ \rho_a = 0.5$) ····· 72
Fig. 4.43 SE of the ensemble mean forecasts ($\rho_e = 0.1,\ \rho_a = 0.5$) ····· 72
Fig. 4.44 SE of the ensemble mean forecasts ($\rho_e = 0.3,\ \rho_a = 0.5$) ····· 73
Fig. 4.45 SE of the ensemble mean forecasts ($\rho_e = 0.5,\ \rho_a = 0.5$) ····· 73
Fig. 4.46 SE of the ensemble mean forecasts ($\rho_e = 0.7,\ \rho_a = 0.5$) ····· 74
Fig. 4.47 SE of the ensemble mean forecasts ($\rho_e = 0.9,\ \rho_a = 0.5$) ····· 74
Fig. 4.48 SE of the ensemble mean forecasts ($\rho_e = 0,\ \rho_a = 0.7$) ····· 76
Fig. 4.49 SE of the ensemble mean forecasts ($\rho_e = 0.1,\ \rho_a = 0.7$) ····· 76
Fig. 4.50 SE of the ensemble mean forecasts ($\rho_e = 0.3,\ \rho_a = 0.7$) ····· 77
Fig. 4.51 SE of the ensemble mean forecasts ($\rho_e = 0.5,\ \rho_a = 0.7$) ····· 77
Fig. 4.52 SE of the ensemble mean forecasts ($\rho_e = 0.7,\ \rho_a = 0.7$) ····· 78
Fig. 4.53 SE of the ensemble mean forecasts ($\rho_e = 0.9,\ \rho_a = 0.7$) ····· 78
Fig. 4.54 SE of the ensemble mean forecasts ($\rho_e = 0,\ \rho_a = 0.9$) ····· 80

Fig. 4.55 SE of the ensemble mean forecasts ($\rho_e = 0.1,\ \rho_a = 0.9$) ····· 80
Fig. 4.56 SE of the ensemble mean forecasts ($\rho_e = 0.3,\ \rho_a = 0.9$) ····· 81
Fig. 4.57 SE of the ensemble mean forecasts ($\rho_e = 0.5,\ \rho_a = 0.9$) ····· 81
Fig. 4.58 SE of the ensemble mean forecasts ($\rho_e = 0.7,\ \rho_a = 0.9$) ····· 82
Fig. 4.59 SE of the ensemble mean forecasts ($\rho_e = 0.9,\ \rho_a = 0.9$) ····· 82
Fig. 4.60 Brier score when $\rho_e =$ 0 ····· 85
Fig. 4.61 Brier score when $\rho_e =$ 0.1 ····· 85
Fig. 4.62 Brier score when $\rho_e =$ 0.3 ····· 86
Fig. 4.63 Brier score when $\rho_e =$ 0.5 ····· 86
Fig. 4.64 Brier score when $\rho_e =$ 0.7 ····· 87
Fig. 4.65 Brier score when $\rho_e =$ 0.9 ····· 87
Fig. 4.66 Summary of Brier score results ····· 88
Fig. 4.67 Summary of Brier score results(by nominal accuracy, ρ_a) ····· 90
Fig. 4.68 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.1$) ····· 92
Fig. 4.69 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.1$) ····· 92
Fig. 4.70 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.1$) ····· 93
Fig. 4.71 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.1$) ····· 93
Fig. 4.72 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.1$) ····· 94
Fig. 4.73 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.1$) ····· 94
Fig. 4.74 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.3$) ····· 96
Fig. 4.75 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.3$) ····· 96
Fig. 4.76 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.3$) ····· 97
Fig. 4.77 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.3$) ····· 97
Fig. 4.78 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.3$) ····· 98
Fig. 4.79 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.3$) ····· 98
Fig. 4.80 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.5$) ····· 100
Fig. 4.81 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.5$) ····· 100
Fig. 4.82 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.5$) ····· 101

Fig. 4.83 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.5$) ····· 101
Fig. 4.84 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.5$) ····· 102
Fig. 4.85 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.5$) ····· 102
Fig. 4.86 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.7$) ····· 104
Fig. 4.87 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.7$) ····· 104
Fig. 4.88 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.7$) ····· 105
Fig. 4.89 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.7$) ····· 105
Fig. 4.90 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.7$) ····· 106
Fig. 4.91 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.7$) ····· 106
Fig. 4.92 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.9$) ····· 108
Fig. 4.93 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.9$) ····· 108
Fig. 4.94 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.9$) ····· 109
Fig. 4.95 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.9$) ····· 109
Fig. 4.96 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.9$) ····· 110
Fig. 4.92 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.9$) ····· 110
Fig. 4.93 Summary of Brier score results ($\rho_a = 0.1$) ····· 115
Fig. 4.94 Summary of Brier score results ($\rho_a = 0.3$) ····· 115
Fig. 4.95 Summary of Brier score results ($\rho_a = 0.5$) ····· 116
Fig. 4.96 Summary of Brier score results ($\rho_a = 0.7$) ····· 116
Fig. 4.97 Summary of Brier score results ($\rho_a = 0.9$) ····· 117
Fig. 4.98 The effective number of scenarios ($\rho_a = 0.1$) ····· 119
Fig. 4.99 The effective number of scenarios ($\rho_a = 0.3$) ····· 119
Fig. 4.100 The effective number of scenarios ($\rho_a = 0.5$) ····· 120
Fig. 4.101 The effective number of scenarios ($\rho_a = 0.9$) ····· 120

기 호

Latin Uppercase

C_v	변동계수 (coefficient of variance)
$\boldsymbol{D_s}$	표준편차로 구성된 대각행렬 (diagonal matrix composed of standard deviation)
$\boldsymbol{I}$	항등행렬 (identity matrix)
$\boldsymbol{R}$	상관 행렬 (correlation matrix)
$\boldsymbol{S}$	공분산 행렬 (covariance matrix)
$\boldsymbol{X}$	앙상블 데이터 행렬 (ensemble data matrix)
$\boldsymbol{Y}$	상관 앙상블 행렬 (correlated ensemble matrix)

Latin Lowercase

f_k	k^{th} 예측 확률 (the probability in the predicted event category k)
f_{ti}	t 시점, i^{th} 구간 예측 확률 (the probability in the predicted event category i at time t)
l	앙상블 시나리오 길이 (length of ensemble scenario)

o_k k^{th} 예측의 관측값 포함여부 2진수 표기
(the binary outcome of the ith event)

o_{ti} t시점, i^{th} 예측 구간의 관측값 포함여부 2진수 표기
(the binary outcome of the event in category i at time t)

p 앙상블 멤버 수 (the number of ensemble members)

$\boldsymbol{x_i}$ i^{th} 앙상블 시나리오 (i^{th} ensemble scenario)

$\overline{\boldsymbol{x}}$ 평균 앙상블 시나리오 (mean ensemble scenario)

$\boldsymbol{y_i}$ i^{th} 상관 앙상블 시나리오 (i^{th} correlated ensemble scenario)

$\boldsymbol{y_o}$ 관측시나리오 (observed scenario)

$\overline{\boldsymbol{y}}$ 평균 상관 앙상블 시나리오 (mean correlated ensemble scenario)

Greek Lowercase

ρ_a 대표 시나리오와 관측시나리오의 상관계수
(nominal accuracy, correlation between representative and observed scenario)

ρ_e 앙상블 시나리오 간 상관계수
(correlation between ensemble scenario)

ρ_{xy} 확률변수 x와 y의 상관계수
correlation between random variable x and y

제 1 장 서 론

1.1 연구의 배경 및 필요성

전 세계적으로 도시화와 기후변화로 인하여 홍수피해가 커지고 있다. 최근 태국은 전례가 없는 홍수피해를 입어 813 명의 인명피해, 이재민 315만 명, 그리고 475억 달러의 피해를 입었다. 또한 2011년에 필리핀에서도 열대 폭풍우 '와시'로 인하여 1,400여명의 인명피해를 입었다(소방방재청, 2011). 우리나라에서도 2011년 6, 7월의 집중호우, 8월 태풍 '무이파'등으로 인하여 피해를 입었으며, 특히 광화문이 침수되고 우면산이 무너지는 사고가 발생하였다. 2011년 재해연보에 따르면 호우로 인하여 발생한 인명피해의 비율이 99% 그리고 재산피해의 비율이 66.4%에 이른다.

이처럼 우리나라는 자연재해로 인한 대부분의 인명피해 및 재산피해가 호우와 태풍으로부터 발생한다. 따라서 호우와 태풍이 야기하는 홍수피해를 줄이기 위하여 많은 예산과 노력이 필요하다. 홍수피해를 줄이기 위한 방법에는 댐, 제방, 저류조 등의 구조물을 건설하는 구조적 대비책과 홍수예·경보 시스템과 같은 비구조적 대비책이 있다. 많은 건설비용과 시간이 필요한 구조적 대비책과는 달리 비구조적 대비책은 홍수의 예·경보를 통하여 인명 피해 방지에 신속한 효과를 볼 수 있다. 홍수 예·경보 시스템은 피해 예상지역 주민들이 대피할 시간을 확보하는데 목적이 있다. 이로 인하여 피해 예상지역의 인명피해 및 재산피해를 줄일 수 있게 되며, 구조적 대비책을 보완하는 역할을 한다.

최근 10년간 홍수 방어와 예보가 중요한 사회 현안이 되면서 홍수예측 기술 향상에 관심이 모아지고 있다(Demeritt et al., 2007; Parker and Fordham, 1996; Pitt, 2007; van Berkom et al., 2007; Cloke and Pappenberger, 2009). 현재 우리나라의 홍수 통제소에서는 중요 지점에

대한 실시간 홍수 예·경보를 시행하고 있다. 홍수 예·경보는 기상청에서 제공하는 MAPLE(McGill Algorithm for Precipitation nowcasting by Lagrangian Extrapolation)과 KLAPS(Korea Local Analysis and Prediction System)등의 강우예측 자료를 입력자료로 사용하여 수위를 예측하는 방식으로 이루어진다. 하지만 이러한 확정적인 단일값만을 제공하는 홍수예측은 확률예측에 비하여 적은 정보를 제공한다는 점에서 한계를 가지고 있다.

전 세계적으로 홍수예측 시스템 분야에서 확정적인 예측보다는 앙상블 예측 시스템(Ensemble Prediction System, EPS)에 대한 연구가 증가하고 있다(Cloke and Pappenberger, 2009). 또한 수문학 분야에서 이를 이용하여 유량예측을 하는 앙상블 유량예측(Ensemble Streamflow Prediction, ESP)에 대한 연구가 활발히 진행되어 왔다. 이러한 앙상블 유량예측은 발생 확률의 범위를 표현하기 위해서 확률예측이 가지는 전체 불확실성을 분석할 필요가 있다. 이를 위하여 많은 불확실성 분석 연구가 진행되어왔다.

앙상블 예측 시스템은 여러 요인들이 가지는 복합적인 작용으로 불확실성을 내포하고 있다. 홍수예측에 있어서 기상 입력변수가 가지는 불확실성이 가장 큰 요인을 제공하는 것으로 알려져 있다. 또한 자료의 보정, 다운스케일링 과정에서 발생하는 불확실성, 수문모형에 사용되는 입력자료가 가지는 시·공간적 불확실성, 홍수예측 시스템이 내포하는 기하학적 불확실성, 강우-유출 모형이 가지는 초기조건 및 매개변수의 불확실성, 그리고 강우-유출 모형이 표현하지 못하는 유출 과정으로 인한 불확실성 등이 있다. 이처럼 강우-유출 과정의 모든 과정에 불확실성에 내포되어 있다. 이를 분석하기 위한 많은 연구들이 이루어지고 있는 실정이다.

앙상블 예측을 수행할 경우 일반적으로 시나리오간 동일한 확률을 가정한다.(Cloke and Pappenberger, 2009). 앙상블 시나리오들이 동등한 확률을 가진다는 것은 시나리오 간 독립성을 가정하고 있음을 의미한다. 서로 독립이라는 가정아래 무작위로 생성된 모든

예측값들의 조합은 어떤 사상에 대한 예측분포의 추정치가 되며 확률론적 예측을 가능케 한다. 그러므로 앙상블 예측은 시나리오 간에 비상관성(uncorrelation)을 요구한다. 하지만 실제 앙상블 예측의 경우 이러한 비상관성이 지켜지기 어려운 실정이다.

우리나라에서는 사용할 수 있는 강우예측 시스템이 한정적이고, 단기 예측에 대하여 시간-지연 앙상블(time-lagged ensemble)을 사용한다. 또한 같은 현상에 대하여 예측을 수행한다는 관점에서 크게 다른 예측 시나리오가 발생하기 어렵기 때문에 앙상블 시나리오 간 상관성이 존재하게 된다. 앙상블 시나리오 간에 독립이 확보되지 못하고 상관성(correlation)이 존재한다면, 예측분포가 왜곡되어 기상 또는 수문예측의 불확실성의 증가되고 신뢰도가 감소할 수 있다 (Stedinger and Kim, 2010). 이러한 앙상블 상관성에 관한 연구는 단기 예측 분야에서 많이 검토되지 못하고 있다.

많은 앙상블 예측 시스템들은 동등한 확률을 가지는 앙상블 멤버로 구성되도록 설계되어 있다(Cloke and Pappenberger, 2009). 따라서 발생 확률의 범위를 표현하기 위하여 적절한 수의 앙상블 시나리오가 설정되어야 한다. 또한 시나리오 수는 예측의 불확실성을 조절하는 중요한 요인이다. 하지만 확률 홍수예측을 위하여 적정한 수의 시나리오를 결정하는 것은 어려운 일이며, 여러 연구가 진행되었으나 적절한 앙상블 시나리오 수에 대한 합의점을 보여주지 못하고 있다. 그리고 시나리오 간 상관성이 주어진 상황을 고려하지 못하고 있다. 따라서 앙상블 시나리오 수와 시나리오 간 상관성이 앙상블 예측 시스템에 미치는 영향을 분석할 필요가 있다.

1.2 연구의 목적

본 연구의 목적은 앙상블 시나리오 간 상관성이 존재할 경우 앙상블 예측 시스템에 미치는 영향을 분석하는 것이다. 앙상블 시나리오 간 상관성이 존재할 때 시나리오 간 독립성 가정이 위배되고 예측의 정확성이 떨어질 수 있다. 이 때 확률예측이 가지는 예측의 정확성을 평가하여, 시나리오 간 상관성과 확률예측 정확성의 관계를 분석하고자 한다. 그리고 높은 정확도와 낮은 정확도의 앙상블 예측을 예상할 수 있는 경우를 나누어 분석하였다. 이와 함께 앙상블 시나리오 개수를 조정하여 상관성 정도에 따라 앙상블 개수가 정확성에 미치는 영향을 분석하였으며, 이를 기반으로 적절한 수의 앙상블을 추정하였다.

1.3 논문의 구성

본 논문의 2장에서는 시·공간 상관관계와 수문예측, 앙상블 유량예측 및 앙상블 불확실성 분석에 관한 연구동향에 대해 살펴보았다. 3장에서는 크게 앙상블을 생성하는 방법과 평가하는 방법을 설명하였다. 앙상블 생성에서는 앙상블 데이터를 정의하고, 상관 난수 생성 방법과 실험 설계를 기술하였다. 앙상블 평가법은 확정론적 예측 평가방법과 확률론적 예측 평가방법으로 구분하였다. 4장에서는 각 예측 평가방법에 대하여 전체 분석 결과와 ρ_a별 분석결과를 나타내었고, 유효 앙상블 시나리오 수를 추정하였다. 마지막으로, 5장에서는 결론 및 향후 연구를 서술하였다.

제 2 장 연구동향

2.1 시·공간 상관관계와 수문 분야

수문 분야에서 시·공간 상관성은 매우 중요한 주제이며, 많은 연구가 진행되어 왔다. Pekarova et al.(2003)은 세계 대하천 연유출량 시계열의 자기 상관성을 분석하여 지역별로 주기성을 연구하였다. Western et al(2004)는 토양수분의 공간상관성과 유역특성의 공간상관성의 패턴을 비교하여 토양수분 분포에 지배적 영향을 미치는 인자를 찾는 방법을 제시하였다. Xu et al(2006)는 양쯔강 유역에 대하여 Penman-Monteith 기준 증발산량과 증발접시 증발량의 시·공간 상관성을 이용하여 두 변수의 공간 분포와 경향을 분석하였다.

수문 시계열을 시간과 공간적으로 분해하기 위하여 분해모형(disaggregation model)을 사용한다. Valencia and Schaake(1973)은 연 단위 시계열을 분기, 월, 일 단위 시계열로 분해하기 위하여 두 시계열 간 상관성을 이용한 분해모형을 제시하였다. Mejia and Rousselle(1976)은 전년도의 시계열과의 상관성을 고려하도록 분해모형을 개선하였다. Lane(1979)는 시계열 자료가 시간에 비해 공간의 크기가 큰 경우 매개변수가 많아지는 단점을 개선한 Lane's condensed model을 제시하였다. Stedinger et al.(1985)는 로그노말 분포를 가정하여 분해모형을 제시하였다.

수문 분야에서 지역 회귀 분석을 수행하는 경우 변수 간 독립성을 가정한다. 하지만 시·공간적으로 상관성이 존재하는 경우 GLS(Generalized least squares) 기법을 이용하여 회귀분석을 수행한다. Stedinger and Tasker(1985)는 수문 자료 간 상관성을 고려하는 GLS가 다른 기법과 비교하였을 때 더 우수한 매개변수 추정을 수행한다고 하였다. Madsen and Rosbjerg(1997)은 공간적 상관성이 있는 48개 유역 홍수 자료에 GLS 회귀분석을 수행하였다. Madsen et al.(2002)는 GLS 회귀모형을

이용하여 확률강우강도곡선(IDF curve)의 지역화 추정방법을 제시하였다.

변수 간의 상관성의 구조를 분석하여 자료의 선형변환을 수행하는 주성분분석(PCA, Principal Components Analysis) 방법이 수문 분야에서 사용한다. Saad and Turgeon(1988) 장기 저수지 운영 최적화를 위하여 저수량에 주성분분석을 수행하였다. Stewart et al.(2004)는 기상 인자와 유량의 주성분분석을 통하여 융설로 인한 유량 변화에 관하여 연구하였다. Najafi et al.(2012)는 PCA(Principle Components Analysis)를 불필요한 변수의 수를 줄이고, 시나리오에 효율적으로 가중치를 적용하는 방법을 제시하였다.

두 변수집단 사이의 상관성을 설명하기 위한 방법으로 정준상관분석(CCA, Canonical Correlation Analysis)이 있다. 이 분석방법은 한 변수집단의 선형결합과 다른집단의 선형결합간의 상관계수가 최대가 되도록 선형결합계수를 산정한다. 수문분야에서는 최적의 관측소 수를 결정하거나 적절한 물리적 변수를 선택하는데 사용될 수 있다. Ourada et al.(2001)은 정준상관분석 방법을 이용하여 지역빈도분석을 수행하였으며, 관측유역과 미관측유역에 대하여 각각 분석 알고리듬을 제시하였다. Tsakiris et al.(2011)은 유역을 균질한 소유역으로 분리하고, 유역 특성 변수와 유량 관련 변수들간의 관계를 찾는데 정준상관분석 방법을 사용하였다.

2.2 앙상블 유량예측

앙상블 유량예측은 미국을 중심으로 가장 활발히 이용되고 있는 확률 유량예측 기법으로, 앙상블 유량예측이 미국 NWS (National Weather Service)의 공식적인 유량예측 기술로 발표된 이후 그 장단점에 대한 논란이 거듭되었으나, NWS의 예측 시스템 현대화 작업의 일환으로 AHPS (Advanced Hydrologic Prediction Service)가 개발되었으며, 현재 Ohio river basin 등의 RFC (River Forecast Center)에서 하천의 수위예측을 위해 사용되고 있다. 현재에는 장기예측에 더해 단기의 유량예측까지 수행할 수 있는 XEFS (Experimental Ensemble Forecast System)에 대한 연구를 수행하고 있다(Day, 1985; Fread, 1985; Smith et al., 1992).

앙상블 유량예측에 관한 연구로는 Day(1985)의 단기간의 기온 및 강우에 대한 예측자료와 현재의 기후 자료를 장기 앙상블 유량예측 기법에 적용하는 개념을 제시하였고, Ingram et al. (1995)는 SVD(Singular Value Decomposition) 방법을 이용하여 과거 강우자료와 예측자료인 QPF(Quantitative Precipitation Forecast)를 가중평균하여 현재시점부터 90일까지를 6시간 간격으로 앙상블 유량예측을 가능하게 하는 방법을 제시하였으며, Schaake et al.(2007)은 NWS에서 앞으로 진행할 연구과제로서 대기-수문모형의 연계운영을 통한 앙상블 유량예측의 개선을 제시하였으며, 이를 계기로 기상예보자료와 앙상블 유량예측을 접목하려는 연구가 활발히 진행되기 시작했다.

최근 앙상블 유량예측에 관하여 가장 활발한 연구동향은 HEPEX (Hydrologic Ensemble Prediction EXperiment)을 들 수 있다. HEPEX는 진보된 확률 수문 예측기법에 연구의 초점을 두고 웹사이트를 통해 전 세계적인 참여가 가능하도록 설립되었으며, 상호간의 연구 공유를 유도하여 신뢰와 확신을 가질 수 있는 앙상블 예

측기법을 얻기 위해 노력하고 있다. 또한 앙상블 유량예측방법이 다른 예측분야에 적용되어 사용될 수 있는지에 대해서도 연구 초점을 두고 있다. 2012년에는 “data assimilation for operational hydrologic forecasting and water resources”라는 주제로 HEPEX 국제 워크샵이 한국에서 개최된 바 있다.

국내에서의 앙상블 유량예측 연구는 Kim et al.(2001)이 건설교통부와 한국건설기술연구원이 매월 초 발간하는 ‘물공급전망’의 개선을 위해 공주지점에 앙상블 유량예측 기법을 적용한 이후 다양한 적용사례를 통해 적용 가능성과 개선방향이 검토되어왔다. 정대일과 김영오(2002)는 충주댐의 유입량에 대하여 앙상블 유량예측 기법을 이용한 확률 유량예측의 결과분석을 통해 앙상블 유량예측의 정확성을 향상시키기 위해서 겨울철에는 강우-유출모형의 모의능력을 개선해야 하며, 여름철에는 강우량예측과 같은 기상예보를 효율적으로 이용해야 한다고 지적한 바 있다. Kim et al.(2004)은 앙상블 예측에서의 강우-유출모형의 문제를 해결하기 위한 또 다른 방법으로써, 물리적 강우-유출모형을 통계기법인 최적선형보정(optimal linear correction) 기법을 이용하여 모의 유량을 보정하는 방안을 결합기법과 함께 제안하였다. 기후변화에 대응하여 과거 강우 시나리오 대신 GCMs 등의 기후모형에서 생성된 기후 앙상블을 수문모형에 입력하여 앙상블 유량예측을 실시하는 방안에 대한 연구도 이루어졌다. 정대일 등(2005)은 일단위 강우-유출모형인 SSARR (Streamflow synthesis and Reservoir Regulation)모형을 이용하여 한강, 낙동강, 섬진강유역의 월 앙상블 유량예측을 적용하였다. 이재경(2005)은 기존 연구에서 다루지 않았던 국내에서 사용가능한 기후예보를 광범위하게 조사하여 체계적으로 정리하였으며, 월간산업기상정보를 이용하여 앙상블 유량예측에 적용할 수 있는 방안을 제시하였다. 강부식 등(2007)도 기상청의 월간 강우예측을 이용하여 앙상블 유량예측에서 생산된 예측유량 시나리오의 가중값(또는 확률)을 조정하는 방안을 개발하여 제시하였으며, 균일한 가중값을 부여한 경우보다 앙상블

유량예측의 정확성이 향상되었음을 증명한 바 있으며, 강태호(2008)는 앙상블 유량예측에 있어서 사전 및 사후처리 기법을 적용하였다. 특히 단일값의 기상자료를 확률적인 값으로 변환하는 사전처리를 적용하였다. 김경진(2011)은 단기 유량예측에서 수치예보자료인 RDAPS(Regional Data Assimilation and Prediction System)를 이용하여 시간지연 앙상블(Time lagged ensemble) 방법을 사용한 후 실시간 보정기법으로 시계열 모형과 인공신경망 방법을 국내에 처음으로 홍수예측에 접목하였다. 홍수기에는 인공신경망 방법보다 시계열 모형방법을 적용하는 것이 적합한 것으로 결과를 제시하였다. 이경태(2012)는 기상청의 기상예보 시스템인 MAPLE (McGill Algorithm for Prediction nowcasting by Lagrangian Extrapolation), KLAPS (Korea Local Analysis and Prediction System) 그리고 UM (Unified Model)의 강우예측결과에 시간지연 앙상블, 다중모형 앙상블을 적용하여 생성한 기상앙상블을 저류함수 모형에 입력한 결과를 분석함으로써 확률론적 홍수예측의 강우 불확실성 분석을 기상 앙상블 종류별, 선행시간별, 공간규모 및 시간단위별로 수행하였다.

2.3 앙상블 불확실성 분석

앙상블 예측 시스템은 각각의 앙상블 시나리오가 같은 발생 확률을 가지고 있으며, 사건 발생의 범위를 표현하기 위하여 적절한 수의 앙상블 시나리오 수를 가지도록 설계되어 있다(Cloke and Pappenberger, 2009). 앙상블 예측 시스템이 가지는 불확실성을 분석하기 위해서는 적절한 수의 시나리오가 결정되어야 한다. 또한 시나리오 간 상관성은 앙상블 예측의 정보 손실을 초래하고 불확실성을 왜곡시킬 수 있다.

Brankovic and Palmer(1997)는 ECMWF(European Centre for Medium-Range Weather Forecasts)를 사용하여 앙상블 스킬 스코어의 분포에 대하여 연구하였다. 이때 앙상블 시나리오와 관측시나리오 간 차이의 상관계수를 이용하여 히스토그램의 계급을 나타내었다. 또한 t-검정을 통하여 최소한의 유용한 앙상블 멤버수를 추정한 결과 약 20 개라는 결론을 내렸다. Buizza et al.(1998)은 앙상블 멤버의 수와 사용한 모델의 해상도를 이용하여 앙상블을 평가하였다. 앙상블 멤버 수를 31개에서 50개로 늘렸을 때 가장 높은 예측 수행능력을 보여주었다. Atger(2001)은 ECMWF EPS의 경우 앙상블 멤버의 수를 51 개에서 21개로 줄여도 예측 능력에 작은 영향만을 끼친다고 하였다. 그러나 Jaun et al.(2008)은 2005년 Rhine 유역 홍수에 대하여 앙상블 예측을 적용하였으며, 시나리오 수 대비 BBS(Brier skill score)를 분석하여 비교적 작은 10 개 정도 앙상블 수의 이점을 주장하였다. 또한 많은 실제 유역에 대한 앙상블 적용 연구(Choi and Beven, 2007; Montanari, 2005; Pappenberger and Beven, 2004)에서 수문학적 모델링 모의에서는 50 개 이상의 많은 수의 앙상블 멤버 수가 필요하다고 하였다.

Stephenson and Doblas-Reyes(2000)은 앙상블 시나리오 데이터를 다변량 데이터로 간주하여 여러 가지 다변량 통계분석 방법을

제안하고 기상 앙상블에 대하여 적용하였다. 또한 공간과 시나리오에 대한 두 종류의 공분산 행렬을 생성하였으며, 시나리오 간 상대적 거리를 분석하였다. Stedinger and Kim(2010)은 앙상블 시나리오에 가중치를 방법으로 pdf-ratio method를 제시하였으며, 앙상블 예측에서 시나리오 간 상관성은 정보의 손실을 유발하고 가중치 구조를 왜곡시킬 수 있다고 주장하였다. Najafi et al.(2012)는 spearman's rank correlation coefficient를 이용하여 기상 인자를 선택한 뒤 유량 앙상블을 생성하였다. 그리고 PCA(Principle Component Analysis)를 사용하여 앙상블에 가중치를 주는 방법을 제안하였다.

제 3 장 연구 방법

3.1 앙상블 생성

본 절에서는 확률 홍수예측을 위하여 상관성(Correlation)이 있는 앙상블 시나리오를 생성하는 방법과 다변량 분석법을 검토하고자 한다. Stephenson 과 Doblas-reyes(2000)는 앙상블을 n 개의 앙상블 멤버와 p 차원 위상 공간으로 구성된 $n \times p$ 데이터 행렬로 나타내어 분석한바 있다. 본 연구에서는 앙상블을 $l \times p$ 데이터 행렬로 간주하여 다변량 분석을 수행하고자 한다. 여기서 l은 앙상블 시나리오의 길이, p는 앙상블 멤버의 수이며 p 개의 변수를 의미한다. 행렬은 볼드체의 알파벳 대문자로 표시할 것이며, 볼드체의 알파벳 소문자는 열벡터를 나타낸다.

3.1.1 다변량 상관관계(Multivariate Correlation)

(1) 2 변수 상관관계

확률예측은 시나리오간의 독립성을 기반으로 한다. 서로 독립이라는 가정아래 무작위로 생성된 모든 예측 값들의 조합은 어떤 사상에 대한 예측분포의 추정치가 되며 확률론적 예측을 가능케 한다. 앙상블 시나리오 간에 독립이 확보되지 못하고 상관성(correlation)이 존재한다면, 예측분포가 왜곡되어 기상 또는 수문예측의 불확실성이 증가되고 신뢰도가 감소할 수 있다. 또한 실제로 앙상블 유량예측을 실시할 경우, 앙상블 시나리오 간 독립성이 보장되지 못하는 것을 확인 할 수 있다. 여기서는 상관관계의 의미를 파악하고, 앙상블 데이터 행렬을 정의하며, 선형 대수적 접근 방법을 이용하여 상

관계수를 나타낼 것이다.

상관관계는 두 변량 사이의 관계를 나타낸다. 한 변량이 증가 또는 감소할 때 다른 변량이 그에 따라 증가 또는 감소한다면 두 변량 사이에는 상관관계가 있다고 한다. 한 변량이 증가할 때 다른 변량 또한 증가하면 양의 상관관계, 한 변량이 감소할 때 다른 변량이 감소하면 음의 상관관계가 있다고 한다. 두 변수간의 상관관계를 식으로 표현하면 식 3.1과 같다.

$$\rho_{xy} = \mathrm{corr}(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sqrt{E(x-\mu_x)^2}\ \sqrt{E(y-\mu_y)^2}} \qquad (3.1)$$

공분산은 x와 y의 크기에 영향을 받기 때문에 공분산으로 두 변수의 관계를 비교하는 것은 문제가 발생할 수 있다. 그러므로 식 3.1에 나타난 바와 같이 공분산을 각각의 표준편차로 나누어 주어 표준화된 공분산을 상관관계라고 한다.

상관관계를 기하학적으로 공간좌표 상에서 살펴보면 두 벡터가 이루는 코사인 값을 나타내게 된다. 3차원 공간에서 두 벡터가 이루는 코사인 각을 벡터로 표현하면 식 3.2와 같다.

$$\begin{aligned}\cos\theta &= \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{b} - (\boldsymbol{b}-\boldsymbol{a})^{\mathrm{T}}(\boldsymbol{b}-\boldsymbol{a})}{2\sqrt{(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b})}} \\ &= \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{b} - (\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} - 2\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b})}{2\sqrt{(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b})}} \\ &= \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}}{\sqrt{(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b})}} \qquad (3.2)\end{aligned}$$

여기서 $\boldsymbol{a}$와 $\boldsymbol{b}$는 3차원 공간에서의 벡터이며 θ는 두 벡터가 이루는 각이다(그림 3.1).

상관관계를 식 3.2와 같은 벡터의 형태로 표현을 할 수 있다. n

개로 구성된 두 개의 변수 $\boldsymbol{x}' = (x_1, x_2, \ldots, x_n)$와 $\boldsymbol{y}' = (y_1, y_2, \ldots, y_n)$의 상관관계는 n 차원 공간에서 두 벡터가 이루고 있는 코사인 값으로 표현될 수 있다(식 3.3).

$$\cos\theta = \frac{(\boldsymbol{x} - \bar{x}\boldsymbol{j})^{\mathrm{T}}(\boldsymbol{y} - \bar{y}\boldsymbol{j})}{\sqrt{[(\boldsymbol{x} - \bar{x}\boldsymbol{j})^{\mathrm{T}}(\boldsymbol{x} - \bar{x}\boldsymbol{j})][(\boldsymbol{y} - \bar{y}\boldsymbol{j})^{\mathrm{T}}(\boldsymbol{y} - \bar{y}\boldsymbol{j})]}} = r_{xy} \tag{3.3}$$

여기서 $\boldsymbol{j}$는 1로 이루어진 $n \times 1$ 벡터를 의미하고, r_{xy}는 표본집단의 상관계수를 나타낸다. 식 3.2에서의 $\boldsymbol{a}$와 $\boldsymbol{b}$를 식 3.3에서의 $\boldsymbol{x} - \bar{x}\boldsymbol{j}$와 $\boldsymbol{y} - \bar{y}\boldsymbol{j}$로 보면 두 식의 형태가 일치한다. 그리고 $\boldsymbol{x} - \bar{x}\boldsymbol{j}$는 $\sum_{i=1}^{n}(x_i - \bar{x})$와 같으므로 식 3.3은 표본집단의 상관계수를 나타내게 된다. $\boldsymbol{x} - \bar{x}\boldsymbol{j}$와 $\boldsymbol{y} - \bar{y}\boldsymbol{j}$가 수직을 이루면 $\boldsymbol{x}$와 $\boldsymbol{y}$의 상관계수가 0이 되고, 같은 방향이면 상관계수가 1, 그리고 반대방향인 180°를 이루면 상관계수가 -1이 된다. 상관계수가 ±1인 경우에 두 변수 $\boldsymbol{x}$와 $\boldsymbol{y}$는 선형관계가 있다고 한다.

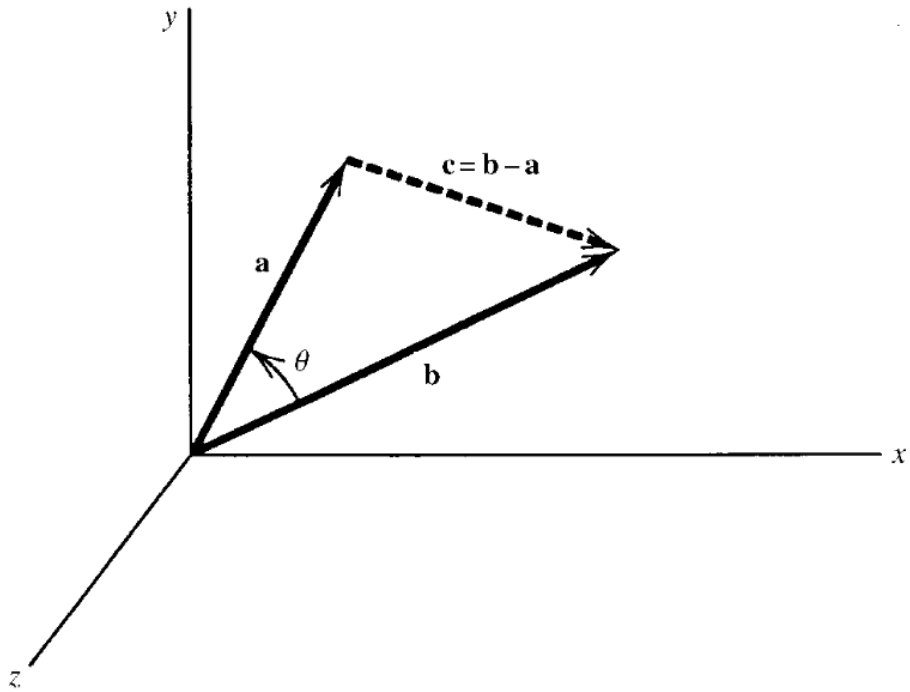


Fig. 3.1 Vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in 3D-space
[출처: Rencher. A.C., 2002]

(2) 다변량 상관관계

앙상블 시나리오간의 상관관계를 구하기 위하여 다변량 상관성 분석이 필요하다. Stephenson and Doblas-Reyes(2000)는 앙상블을 다변량 데이터로 간주하여, 다변량 통계분석을 실시하였다. 본 연구에서는 앙상블을 $l \times p$ 데이터 행렬로 간주하여 p 개의 변수간의 상관관계를 구하는 다변량 분석을 수행할 것이다.

$$\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ \ldots\ , \boldsymbol{x_j}, \ \ldots\ , \boldsymbol{x_p}) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{l1} & x_{l2} & \cdots & x_{lj} & \cdots & x_{lp} \end{pmatrix} \tag{3.4}$$

여기서 $l \times p$ 앙상블 데이터 행렬을 살펴보면, p 개의 각각의 앙상블 시나리오와 l 개의 유량 예측값 또는 시나리오 길이로 구성되어 있다(식 3.4). 그리고 $\boldsymbol{x_j^{\mathrm{T}}} = (x_{1j},\ x_{2j},\ \ldots\ , x_{lj})$와 $\boldsymbol{x_i^{\mathrm{T}}} = (x_{i1},\ x_{i2},\ \ldots\ , x_{ip})$로 행렬 $\boldsymbol{X}$의 행과 열을 정의하였다. 즉, $\boldsymbol{x_j}$는 각각의 앙상블 시나리오를 의미하며, $\boldsymbol{x_i}$는 그 시점에서의 앙상블 예측값들을 의미하게 된다. 또한 평균 벡터 $\overline{\boldsymbol{x}}$를 정의하면 식 3.5와 같다.

$$\overline{\boldsymbol{x}} = \frac{1}{l}\sum_{i=1}^{l}\boldsymbol{x_i} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_p} \end{pmatrix} \tag{3.5}$$

여기서 $\overline{x}_1$은 $(1/l)\sum_{i=1}^{l} x_{i1}$을 나타낸다. 그러므로 $\overline{\boldsymbol{x}}$는 각각 p개의 시나리오들의 평균으로 이루어진 벡터이다. 그리고 이것을 행렬표기법으

로 표현하면 식 3.6과 같다.

$$\bar{\boldsymbol{x}} = \frac{1}{l}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{j} \tag{3.6}$$

상관관계 행렬을 구하기 위해서는 우선 공분산 행렬을 얻어야 한다. p 개의 시나리오간의 공분산을 구해야 하기 때문에 공분산 행렬은 $p \times p$ 행렬로 구성되어 있다. $p \times p$ 공분산 행렬을 정의하면 식 3.7 과 같다.

$$\boldsymbol{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2k} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jk} & \cdots & s_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pk} & \cdots & s_{pp} \end{pmatrix} \tag{3.7}$$

공분산은 두 변수간의 관계를 나타내며, 식 3.1에서 알 수 있듯이 순서에 무관하므로 $s_{jk} = s_{kj}$ 가 된다. 따라서 공분산 행렬 $\boldsymbol{S}$는 주대각선에 대칭인 두 원소가 같은 대칭행렬이다. 주대각선 위의 성분인 s_{jj}와 주대각선 밖의 성분인 s_{jk}는 식 3.8과 식 3.9와 같다.

$$\begin{aligned} s_{jj} &= s_j^2 = \frac{1}{l-1}\sum_{i=1}^{l}(x_{ij} - \overline{x_j}) \\ &= \frac{1}{l-1}\left(\sum_{i=1}^{l} x_{ij} - l\overline{x_j}^2\right) \end{aligned} \tag{3.8}$$

$$\begin{aligned} s_{jk} &= \frac{1}{l-1}\sum_{i=1}^{l}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k}) \\ &= \frac{1}{l-1}\left(\sum_{i=1}^{l} x_{ij}x_{ik} - l\overline{x_j}\,\overline{x_k}\right) \end{aligned} \tag{3.9}$$

주대각선상의 성분이 s_{jj}는 식 3.8에서 알 수 있듯이 j 번째 시나리오의 분산을 의미한다. 그리고 식 3.9에서 보듯이 s_{jk}는 j 번째 시나리오와 k 번째 시나리오의 공분산을 나타낸다. 공분산 행렬을 벡터로 표현하면 식 3.10과 같다.

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{l-1}\sum_{i=1}^{l}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})^{\mathrm{T}} \\ &= \frac{1}{l-1}\left(\sum_{i=1}^{l}\boldsymbol{x_i}\boldsymbol{x_i}^{\mathrm{T}} - l\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^{\mathrm{T}}\right) \end{aligned} \tag{3.10}$$

여기서 $\boldsymbol{x_i}-\overline{\boldsymbol{x}}$는 $p \times 1$ 벡터이므로 $(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})^{\mathrm{T}}$는 p × p 행렬이다. i를 1부터 l까지 더하면 식 3.8과 식 3.9로 구성된 공분산 행렬이 된다. 다시 위의 식을 시그마를 사용하지 않고 데이터 행렬을 이용하여 나타내면 식 3.11과 같다.

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{l-1}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{X}^{\mathrm{T}}\left(\frac{1}{l}\boldsymbol{J}\right)\boldsymbol{X}\right] \\ &= \frac{1}{l-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{I} - \frac{1}{l}\boldsymbol{J}\right)\boldsymbol{X} \end{aligned} \tag{3.11}$$

여기서 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$는 식 3.10에서 $\sum_{i=1}^{l}\boldsymbol{x_i}\boldsymbol{x_i}^{\mathrm{T}}$와 같다. 그리고 $\overline{\boldsymbol{x}}$를 행렬로 표현한 식 3.6을 이용하면 $\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^{\mathrm{T}}$는 $\boldsymbol{X}^{\mathrm{T}}\left(\frac{1}{l^2}\boldsymbol{J}\right)\boldsymbol{X}$가 된다. 식 3.11에서 $\boldsymbol{J}$는 성분이 모두 1로 이루어진 $p \times p$ 행렬이다.

식 3.1을 살펴보면 상관관계는 공분산을 두 변수의 표준편차로 나누어져 있음을 알 수 있다. 이 관계를 이용하여 상관관계 행렬을 구할 수 있다. 우선 주대각선 상위의 성분은 표준편차로 이루어져 있고, 나머지 성분은 모두 0인 대각 행렬을 정의하면 식 3.12와 같다.

$$\boldsymbol{D_s} = \begin{pmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{s_{jk}} & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sqrt{s_{pp}} \end{pmatrix} \tag{3.12}$$

식 3.1에서 나타난 공분산과 상관관계의 관계와 대각 행렬 $\boldsymbol{D_s}$, 그리고 공분산 행렬 $\boldsymbol{S}$를 이용하여 상관관계 행렬 $\boldsymbol{R}$을 나타내면 식 3.13과 같다.

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{D_s^{-1}} \boldsymbol{S} \boldsymbol{D_s^{-1}} \\ &= (\rho_{jk}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2k} & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ \rho_{j1} & \rho_{j2} & \cdots & 1 & \cdots & \rho_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pk} & \cdots & 1 \end{pmatrix} \end{aligned} \tag{3.13}$$

상관관계 행렬은 공분산 행렬과 마찬가지로 $p \times p$ 행렬로 구성되어 있다. 그리고 식 3.13에서 나타나듯이 주대각선에 대칭인 성분이 같은 대칭행렬로 구성되어 있으며, 주대각선의 성분은 모두 1로 구성되어 있다. ρ_{jk}는 j 번째 시나리오와 k 번째 시나리오의 상관관계를 나타낸다. 따라서 주대각선의 성분인 ρ_{jj}는 모두 1임을 알 수 있다.

3.1.2 상관 난수 생성(Correlated Random Number Generation)

(1) 2 변수 상관 난수 생성

상관 앙상블 시나리오의 효과를 확인하기 위해서는 임의의 상관계수를 갖는 시나리오를 생성을 해야 한다. 상관 난수 생성은 이변수 생성부터 시작되었다. Hoffan(1959)는 상관계수가 0인 확률 변수 X와 Z가 있다고 가정하고, 확률변수 X와 임의의 상관계수 ρ_{xy}를 갖는 확률변수 Y를 X와 Z의 선형적인 관계로 표현하였다(식 3.14).

$$Y_i = X_i + bZ_i \tag{3.14}$$

여기서 b는 상수이며, 확률변수 X와 Z의 평균을 0이라고 가정하였다. 그리고 확률변수 X와 Y의 임의의 상관계수를 나타내면 다음 식 3.15와 같다.

$$\begin{aligned}\rho_{xy} &= \sum_{i=1}^{N} \frac{(x_i)(x_i + bz_i)}{N\sigma_x \sigma_{x+bz}} \\ &= \frac{\sigma_x^2 + b\rho_{xz}\sigma_x\sigma_z}{\sigma_x\sigma_{x+bz}}\end{aligned} \tag{3.15}$$

처음 가정에서 ρ_{xz}는 0이므로 식 3.15는 식 3.16과 같다.

$$\rho_{xy} = \frac{\sigma_x}{\sigma_{x+bz}} \tag{3.16}$$

또한 $X+bZ$의 분산 σ_{x+bz}는 $\sigma_x^2 + b^2\sigma_z^2$과 같기 때문에 이것을 3.16에 대입하면 식 3.17과 같다.

$$\rho_{xy}^2 = \frac{\sigma_x^2}{\sigma_x^2 + b^2\sigma_z^2} \tag{3.17}$$

임의의 상관계수 ρ_{xz}와 확률변수 X와 Z의 분산 σ_x^2와 σ_z^2를 알고 있으므로 상수 b를 구할 수 있다(식 3.18)

$$b = \frac{\sigma_x}{\sigma_z}\sqrt{\frac{1-\rho_{xy}^2}{\rho_{xy}^2}} \tag{3.18}$$

따라서 식 3.19를 사용하면 원하는 평균, 분산, 그리고 확률변수 X와의 상관계수를 갖는 확률변수 Y를 구할 수 있다.

$$Y_i = \frac{\sigma_{yA}}{\sigma_y}(X_i + bZ_i) + C \tag{3.19}$$

여기서 σ_y는 앞서 구한 σ_{x+bz}이며, σ_{yA}는 원하는 확률변수 Y의 표준편차이다. 확률변수 X와 Z의 평균을 0으로 가정하였기 때문에 확률변수 Y의 평균은 C가 되게 된다.

확률변수 X와 Z가 각각 표준정규분포를 따르면 평균은 각각 0이고 표준편차 $\sigma_x = \sigma_z = 1$이 된다. 이 경우에 임의의 상관계수 ρ_{xy}를 갖는 확률변수 Y를 X와 Z의 선형적인 관계로 표현하면 식 3.20과 같다.

$$Y = \rho_{xy}X + \sqrt{1-\rho_{xy}^2}\,Z \tag{3.20}$$

여기서 X와 Z를 각각 표준정규분포를 따르도록 난수 생성을 한 뒤에 식 3.20에 따라 Y를 구하면 임의의 상관계수 ρ_{xy}를 갖도록 할 수 있다. 그리고 Y 또한 식 3.20의 관계에 의해서 표준정규분포를 따르게 된다.

(2) 콜레스키 분해(Cholesky Decomposition)

콜레스키 분해는 다변량 상관 난수 생성을 위하여 필요한 행렬 분해법이다. 이것은 양정치 행렬(positive definite matrix)를 상삼각 행렬(upper triangular matrix)의 행렬곱으로 표현하는 방법이다(식 3.21)

$$\boldsymbol{A} = \boldsymbol{T}^{\mathrm{T}}\boldsymbol{T} \tag{3.21}$$

여기서 행렬 $\boldsymbol{A}$는 대칭 행렬로 다음 식 3.22를 만족하면 양정치 행렬이라고 한다.

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} > 0 \qquad \forall \boldsymbol{x} \neq 0 \tag{3.22}$$

또한 식 3.21의 $\boldsymbol{T}$는 상삼각 행렬이며, 정사각 행렬의 특수한 경우로 주대각선 아래쪽 성분들의 값이 모두 0인 경우를 말한다. 실제 앙상블을 생성을 하면 시나리오 간 상관계수가 모두 양의 실수로 나타나기 때문에 앙상블 시나리오 상관 행렬은 식 3.22의 조건을 만족하여 양정치 행렬이 된다.

양정치 행렬 $\boldsymbol{A}$를 상삼각 행렬 $\boldsymbol{T}$로 콜레스키 분해를 하면 다음 식 3.23과 같다. 이 식에서 $\boldsymbol{A} = (a_{ij})$ 그리고 $\boldsymbol{T} = (t_{ij})$로 나타내었다.

$$t_{11} = \sqrt{a_{11}}, \quad t_{1j} = \frac{a_{1j}}{t_{11}} \qquad 2 \leqq j \leqq n,$$

$$t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t_{ki}^2} \qquad 2 \leqq i \leqq n$$

$$t_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} t_{ki}t_{kj}}{t_{ii}} \qquad 2 \leqq i < j \leqq n,$$

$$t_{ij} = 0 \qquad 1 \leqq j < i \leqq n. \qquad (3.23)$$

앞서 2 변수 상관 난수 생성 파트에서 사용하였던 식 3.20에도 콜레스키 분해법을 사용할 수 있다. 앞에서는 확률변수 X, Y, 그리고 Z로 표현하였지만 선형대수학적인 방법을 사용하기 위해서 각각 $n \times 1$ 벡터 $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$를 사용하였다. 우선 $\boldsymbol{x}$와 $\boldsymbol{y}$의 상관 행렬은 다음 식 3.24와 같다.

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{pmatrix} \qquad (3.24)$$

여기서 $\boldsymbol{x}$와 $\boldsymbol{y}$의 상관 행렬 $\boldsymbol{R}$은 식 3.22를 만족하는 양정치 행렬로 가정한다. 그리고 상관 행렬 $\boldsymbol{R}$을 콜레스키 분해법을 사용하여 분해하면 다음 식 3.25와 같다.

$$\boldsymbol{T} = \begin{pmatrix} 1 & \rho_{xy} \\ 0 & \sqrt{1-\rho_{xy}^2} \end{pmatrix} \qquad (3.25)$$

상삼각 행렬 $\boldsymbol{T}$로 분해된 것을 확인할 수 있으며, 식 3.20에 있는 확률변수 X와 Z 앞에 있는 계수가 성분으로 있음을 볼 수 있다. 식 3.26을 이용하면 $\boldsymbol{x}$와 상관계수 ρ_{xy}를 갖는 시나리오 $\boldsymbol{y}$를 생성할 수 있음을 알 수 있다.

$$(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{z}) \begin{pmatrix} 1 & \rho_{xy} \\ 0 & \sqrt{1-\rho_{xy}^2} \end{pmatrix} \qquad (3.26)$$

(3) 다변량 상관 난수 생성

다변량 상관 난수 생성은 Kaiser and Dickman(1962)이 처음 정규 분포를 따르는 모집단으로부터 상관 데이터를 생성하는 방법을 제안하였다. 앞서 서술된 Hoffan(1959)의 2 변수 상관 난수 생성 방법에서 상관된 변수의 개수를 확장하였다. 다변량 상관 난수의 경우에도 식 3.26의 방법과 마찬가지로 상관 행렬을 분해한 행렬과 다변량 데이터 행렬의 곱으로 표현하였다. 다변량 상관 난수 생성 방법은 다변수 정규분포의 선형변환 과정에서 찾을 수 있다. 다변수 정규분포는 다음 식 3.27과 같다.

$$g(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/2} \quad , \quad N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.27}$$

여기서 p는 변수의 개수를 의미한다. 따라서 $\boldsymbol{y}$는 $p \times 1$ 형태의 벡터로 이루어져 있음을 알 수 있다. 이 분포는 평균이 $\boldsymbol{\mu}$이고 공분산 행렬 $\boldsymbol{\Sigma}$를 따른다. 그리고 자연상수 위의 지수 $(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$는 $\boldsymbol{y}$의 $\boldsymbol{\mu}$로부터의 거리를 나타내며, 마하라노비스의 거리(Mahalanobis distance)라고 불린다.

일변량 정규분포(univariate normal distribution)의 경우 평균과 분산의 단일 값으로 표현된다. 반면에 다변수 정규분포의 경우 평균은 벡터로 표현되고 분산은 공분산 행렬로 나타나게 된다. 다변수 분포의 선형변환 과정에서 나타나는 공분산 행렬의 변화를 이용하여 다변량 상관 난수를 생성할 수 있다. 식 3.28은 다변수 정규분포의 선형변환을 나타낸다.

$$\begin{aligned} \boldsymbol{y} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{A}\boldsymbol{y} &\sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}) \end{aligned} \tag{3.28}$$

여기서 $\boldsymbol{A}$는 $q \times p$ 행렬이며, 선형 변환 후 $\boldsymbol{Ay}$는 q 개의 변수를 가지고 평균 $\boldsymbol{A\mu}$, 공분산 $\boldsymbol{A\Sigma A}^{\mathrm{T}}$를 따르는 다변수 정규분포가 된다. 이 성질을 이용하면 임의의 공분산을 가지는 다변량 상관 난수를 생성할 수 있다(식 3.29).

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{F}^{\mathrm{T}}\boldsymbol{Z} \\ \boldsymbol{Z} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{W} &\sim N_p(\boldsymbol{F\mu}, \boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Sigma F}) \end{aligned} \qquad (3.29)$$

여기서 $\boldsymbol{Z}$는 $p \times n$ 형태의 행렬로 p 개의 변수와 n 개의 관측값으로 이루어져 있다. 그러므로 행렬 $\boldsymbol{Z}$는 앞서 식 3.4의 앙상블 데이터 행렬 $\boldsymbol{X}$의 전치행렬이다. $\boldsymbol{Z}$는 평균 $\boldsymbol{\mu}$와 공분산 $\boldsymbol{\Sigma}$를 따른다.

Z의 공분산 행렬을 $\boldsymbol{\Sigma} = \boldsymbol{I}$ 라고 가정한다. 여기서 $\boldsymbol{I}$는 $p \times p$ 단위행렬이다. 그리고 $\boldsymbol{F}$가 임의의 공분산 행렬 $\boldsymbol{S}$를 콜레스키 분해로 얻은 행렬이라고 하면 $\boldsymbol{W}$의 공분산은 $\boldsymbol{S}$가 된다. 그리고 행렬 $\boldsymbol{W}$와 $\boldsymbol{Z}$를 전치시켜서 다시 기술하면 식 3.30과 같다.

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{XF}, \quad \boldsymbol{S} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} \\ \boldsymbol{X} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{I}) \\ \boldsymbol{Y} &\sim N_p(\boldsymbol{F\mu}, \boldsymbol{S}) \end{aligned} \qquad (3.30)$$

여기서 행렬 $\boldsymbol{X}$는 $l \times p$ 앙상블 데이터 행렬이다. $\boldsymbol{S}$는 $p \times p$ 공분산 행렬이므로 $\boldsymbol{F}$ 또한 $p \times p$ 행렬이 된다. 따라서 행렬 $\boldsymbol{X}$와 마찬가지로 새로 생성된 행렬 $\boldsymbol{Y}$는 $l \times p$ 행렬이 된다. 그리고 공분산 행렬은 대칭행렬이므로 전치 후에도 같은 값을 가지므로 식 3.29와 식 3.30의 $\boldsymbol{F}$가 서로 같음을 알 수 있다. 또한 $\boldsymbol{X}$의 평균을 $\boldsymbol{0}$으로 하면 $\boldsymbol{Y}$의 평균도 $\boldsymbol{0}$이 되므로 같은 평균을 가지며 임의의 공분산을 가지는 데이터를 생성하게 된다.

공분산 행렬의 주대각선의 성분, 즉 각각 시나리오의 분산이 1

이라면 식 3.12의 대각 행렬 D_S가 단위행렬 I가 되어 공분산 행렬 S와 상관관계 행렬 R이 같게 된다. 따라서 평균이 0이고 공분산이 단위행렬 I인 행렬을 생성한 뒤 임의의 상관관계 행렬을 분해하여 곱하면 평균이 0이고 임의의 상관관계 행렬을 가지는 행렬을 생성할 수 있다.

3.1.3 상관 앙상블 생성 및 관측시나리오 생성

몬테카를로 시뮬레이션은 변수를 조작하고 통계적 특성의 변화의 영향을 측정하기 위하여 매개변수를 알고 있는 모집단으로부터 난수 표본 집단을 생성하는 기법이다. 즉 변수간의 관계가 명확하지 못하여 정확한 해를 구할 수 없는 불확실한 상황 하에서의 확률적 시스템의 모의실험에 이용되는 절차를 말한다. 본 연구에서는 앞서 언급한 다변량 상관 난수를 생성하여 앙상블 시나리오 간 독립성이 지켜지지 못하고 상관관계가 존재하는 경우에 대하여 모의실험을 수행할 것이다.

본 연구에서는 10 가지 앙상블 멤버 수를 가지는 앙상블을 생성하여 실험을 할 것이다. 그 수는 3, 5, 7, 9, 12, 15, 20, 30, 50, 그리고 100개이다. 또한 시나리오의 길이 l은 1000으로 설정하였다. Stephenson and Doblas-Reyes(2000)은 앙상블 예측을 위하여 미래 전체의 확률분포에 대한 이해보다는 미래 예측값의 올바른 추정과 초기 조건에 대한 예측의 민감도를 중요시하여 평균과 공분산 추정에 정규분포(Gaussian distribution)가 유용하다고 기술하였다. 따라서 각각의 앙상블 시나리오는 정규분포를 따른다고 가정하고, 분석의 편리함을 위하여 평균이 0이고, 표준편차가 1인 표준정규분포를 사용하였다. 그리고 시나리오 간의 상관계수(ρ_e)는 0, 0.1, 0.3, 0.5, 0.7, 그리고 0.9로 6 가지를 사용하였다.

상관 앙상블을 생성하는 방법을 정리하면 다음과 같다. 표준정규분포를 따르며 길이가 1000인 시나리오들을 10 가지 종류의 멤버 수를 가지도록 생성한다(matrix $\boldsymbol{X}$). 그리고 6가지 종류의 시나리오 상관 계수를 나타내는 상관관계 행렬 $\boldsymbol{R}$을 만든다. 상관관계 행렬 $\boldsymbol{R}$을 콜레스키 분해법을 사용하여 행렬 $\boldsymbol{T}$를 생성한다. 마지막으로 행렬 $\boldsymbol{X}$를 행렬 $\boldsymbol{T}$를 이용하여 선형변환을 하여 상관 앙상블을 얻는다(행렬 $\boldsymbol{Y}$).

표준정규분포 사용의 가능성을 검증하기 위하여 모의과정에서

변동계수(coefficient of variation)를 사용하였다. McMahon and Mein(1986)은 126개 강들의 연홍수량을 조사하여 거의 정규분포를 따르며 변동계수의 중앙값이 0.25라고 분석하였다. 그리고 McMahon et al.(2007)은 세계 1221개 강들의 연홍수량에 대하여 통계분석을 하였다. 감마 분포(Gamma distribution)가 연홍수량을 표현하는데 적합하다고 하였으며, 변동계수의 최소값은 0.0619, 최대값은 2.97, 그리고 중앙값은 0.314라고 하였다(그림 3.2). 따라서 변동계수를 0.1, 0.5, 그리고 1로 설정하여 상관앙상블을 생성하여 모의를 진행하였다. 평균 0 과 표준편차 1을 따르는 상관앙상블 $\boldsymbol{Y}$를 생성한 뒤 누적확률로 변환하였다. 이 누적확률을 평균 1 과 표준편차 C_v를 따르는 누적분포 함수에 적용하여 확률 변수를 구하였다. 이 확률변수는 변동계수를 조절한 새로운 상관앙상블이며 평균 1이며 표준편차가 C_v인 분포를 따르게 된다. 실제 연홍수량의 변동계수를 사용하여 얻은 결과의 확정적 평가지표와 표준정규분포를 사용하여 얻은 확정적 평가 지표를 비교하여 표준정규분포의 사용 가능여부를 판단하는데 목적이 있다.

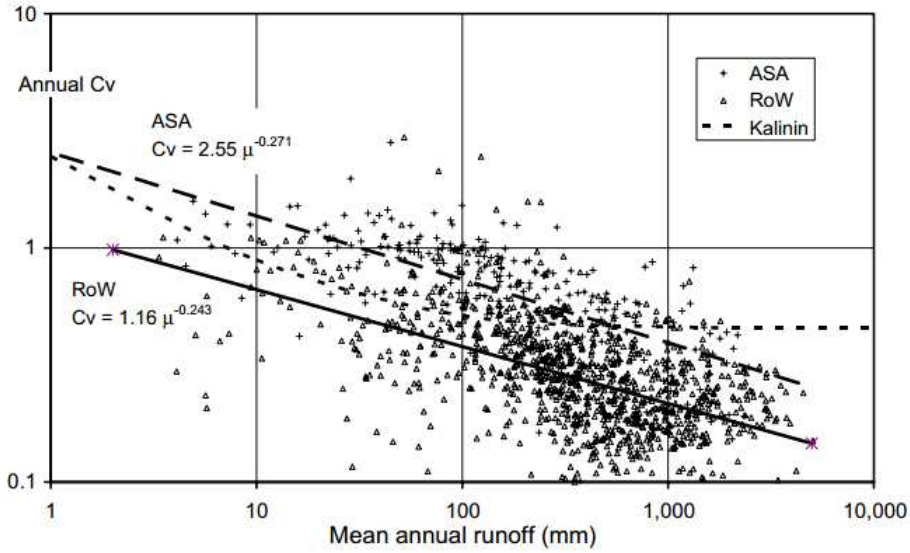


Fig. 3.2 Coefficient of variation of annual flows vs mean annual runoff

[출처: McMahon et al. (2007)]

상관성이 앙상블에 미치는 영향을 분석하기 위해서는 앙상블 생성뿐만 아니라 모의결과를 평가하기 위해서 관측값 또한 생성이 필요하다. 가능한 모든 상황에서 모의를 하기위해서 예측의 정확도별로 효과를 분석하는 방법을 선택하였다. 앙상블 예측이 어느 정도의 정확도를 가진 예측을 수행하는가에 따라 앙상블 시나리오 간 상관관계가 미치는 영향이 다를 수 있다. 그러므로 앙상블 시나리오부터 역으로 정확도를 다르게 가지는 관측값을 생성하여 상관관계와 앙상블의 개수가 미치는 영향을 분석하였다. 식 3.4의 앙상블 데이터 행렬을 식 3.30을 이용하여 임의의 상관관계 행렬을 따르는 앙상블 데이터 행렬 $\boldsymbol{Y}$를 만들면 다음 식 3.31과 같다.

$$\begin{aligned}\boldsymbol{Y} &= (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_j, \ldots, \boldsymbol{y}_p) \\ &= (\boldsymbol{x}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_j, \ldots, \boldsymbol{y}_p)\end{aligned} \tag{3.31}$$

여기서 앙상블 데이터 행렬 $\boldsymbol{Y}$의 첫 번째 열인 $\boldsymbol{y}_1$이 초기의 앙상블 데이터 행렬 $\boldsymbol{X}$의 첫 번째 열인 $\boldsymbol{x}_1$이 같다. 각 시나리오의 표준편차를 1로 가정하였기 때문에 식 3.23에 의하여 분해된 행렬 $\boldsymbol{F}$의 성분 t_{11}이 1이 되고 첫 번째 열의 나머지 성분들이 0이 되어 행렬 $\boldsymbol{X}$와 행렬 $\boldsymbol{Y}$의 첫 번째 열이 같아진다. 따라서 시나리오 $\boldsymbol{x}_1$은 처음 행렬 $\boldsymbol{X}$를 생성한 이후로 변하지 않으며, 상관 앙상블 $\boldsymbol{Y}$를 생성할 시 나머지 모든 시나리오에 영향을 주기 때문에 대표성을 가지게 된다. 그러므로 관측시나리오를 만들기 위하여 시나리오 $\boldsymbol{x}_1$를 사용하였다.

시나리오 $\boldsymbol{x}_1$와 임의의 상관계수(ρ_a, nominal accuracy)를 가지는 시나리오를 하나 생성하여 이를 관측값으로 사용하였다. 각 시나리오가 표준정규분포를 따르도록 하였으므로 시나리오 $\boldsymbol{x}_1$과의 상관계수를 정확도를 나타내는 지표로 사용하였다. 여기서 상관계수는 0.1, 0.3, 0.5, 0.7, 0.9을 사용하였다. 이와 같이 기술한 방법으로 앙상블 데이터 행렬을 생성하고 관측값을 두 가지 방법으로 생성하여 앙상블 시나리오의 개수와 시나리오 간 상관관계가 앙상블 예측에

미치는 영향을 분석하였다. 그리고 각각의 시나리오 개수, 시나리오 간 상관관계와 관측값 생성 방법의 종류별로 1000번의 모의를 시행하였다. 앙상블 생성 및 모의 방법을 요약하면 그림 3.2와 같다.

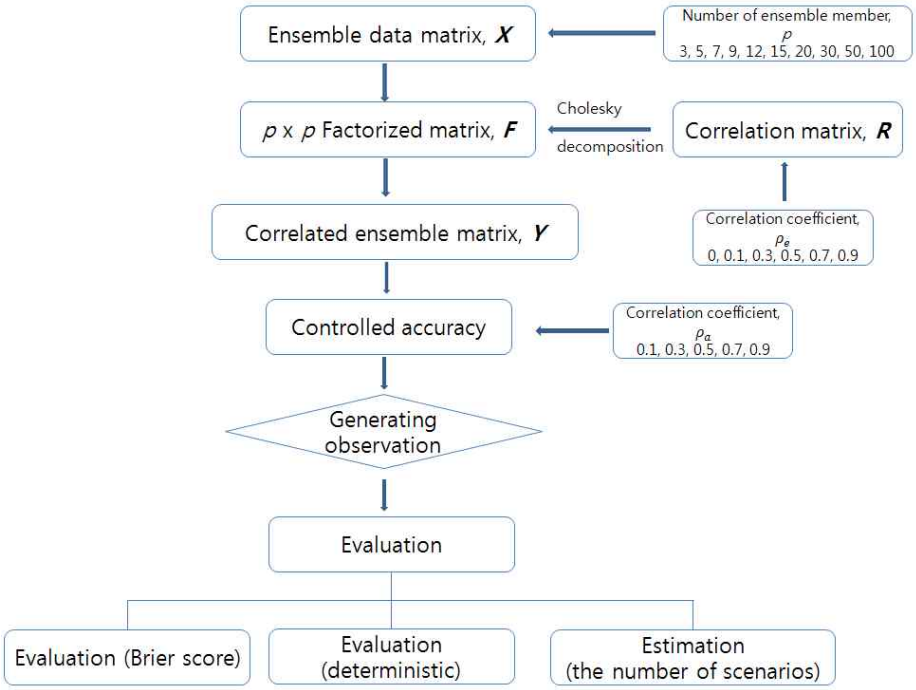


Fig. 3.3 Procedure of study

3.2 앙상블 평가방법

3.2.1 확정적 예측 평가방법

확정론적 예측이란 하나의 단일값으로 미래를 예측을 하는 방법이다. 이러한 단일한 예측값을 평가하기 위하여 주로 세 가지 평가지표를 사용해왔다. 강태호(2008)와 이경태(2012)는 확정론적 예측을 평가하기 위하여 계통적 오차(systematic error)를 보여줄 수 있는 Bias와 무작위 오차(random error)를 보여줄 수 있는 $\sqrt{\mathrm{Variance}}$, 그리고 앞의 두 오차를 함께 보여주는 지표인 RMSE(Root Mean Squared Error)를 사용하였다.

Stephenson and Doblas-Reyes(2000)은 앙상블은 다변량 위상공간에서 관측값과의 거리를 나타내기 위하여 유클리드 노름(Euclidean norm)을 사용하였다(식 3.32).

$$\|\boldsymbol{x}\|_G^2 = \frac{1}{l}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x} = \frac{1}{l}\sum_{i,j=1}^{l} x_i G_{ij} x_j \qquad (3.32)$$

여기서 $\|\boldsymbol{x}\|_G^2$는 벡터 $\boldsymbol{x}$의 제곱 거리(squared length)이다. 그리고 $\boldsymbol{G}$는 다변량 가중 행렬(multivariate weighting matrix)이며 가중치를 주지 않았으므로 단위행렬 $\boldsymbol{I}$를 사용하였다.

본 연구에서는 Stephenson and Doblas-Reyes(2000)이 사용한 평가지표를 사용하였다. 이 평가지표는 강태호(2008)와 이경태(2012)가 사용한 지표와 내용적으로 동일하나 앙상블을 다변량 데이터로 분석하는데 있어서 더 직관적인 수식을 제공한다. 첫 번째 지표는 위상공간에서 앙상블 시나리오들이 퍼져있는 정도를 나타내는 mean squared dispersion, D^2, 이다. 이 지표는 앞서 언급한 $\sqrt{\mathrm{Variance}}$의 제곱의 형태이며 무작위 오차를 보여준다. mean squared dispersion

은 다음 식 3.33과 같다.

$$D^2 = \frac{1}{p}\sum_{i=1}^{p}\|\boldsymbol{y_i} - \overline{\boldsymbol{y}}\|^2 = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{l}\sum_{j=1}^{l} z_{ij}^2 \qquad (3.33)$$

여기서 $\boldsymbol{y_i}$는 앙상블 시나리오이고 $\overline{\boldsymbol{y}}$은 앙상블의 평균시나리오이다. 그리고 $\boldsymbol{z_i} = \boldsymbol{y_i} - \overline{\boldsymbol{y}}$이며 각 앙상블 시나리오와 평균시나리오의 차이를 나타낸다. 따라서 D^2은 각 앙상블 시나리오가 평균 시나리오로부터 떨어져있는 거리 제곱의 평균을 나타낸다. 그러므로 mean squared dispersion이 앙상블 시나리오가 평균으로부터 퍼져있는 정도를 나타내게 되는 것이다.

두 번째 지표는 앙상블 평균이 관측값에 얼마나 가까운지를 나타내는 지표인 SE(Squared Error)이다. 이 지표 또한 앞서 언급한 bias에 대응되는 지표이다(식 3.34).

$$\text{Squared error} = \|\overline{\boldsymbol{y}} - \boldsymbol{y_o}\|^2 \qquad (3.34)$$

여기서 $\boldsymbol{y_o}$는 관측시나리오이다. 앙상블 평균 예측(ensemble mean forecasts)를 할 경우의 오차를 의미한다.

마지막 지표는 전체 앙상블 예측의 MSE(Mean Squared Error)를 나타내는 S^2이다. 이 지표는 D^2과 SE의 합으로 나타낼 수 있으며, 식으로 나타내면 다음과 같다(식 3.35).

$$S^2 = \frac{1}{p}\sum_{i=1}^{p}\|\boldsymbol{y_i} - \boldsymbol{y_o}\|^2 = D^2 + \|\boldsymbol{y} - \boldsymbol{y_o}\|^2 \qquad (3.35)$$

S^2은 각각의 앙상블과 관측시나리오의 거리 제곱의 평균을 나타낸다. 그러므로 앙상블 전체가 가지는 오차를 나타낸다. 식 3.35에 나타나 있듯이 S^2은 나머지 두 지표의 합으로 표현되는데 그 증명은

식 3.36에 나타나있다.

$$
\begin{aligned}
S^2 &= \frac{1}{p}\sum_{i=1}^{p}\|\boldsymbol{y_i}-\boldsymbol{y_o}\|^2 \\
&= \frac{1}{p}\sum_{i=1}^{p}\|\bar{\boldsymbol{y}}-\boldsymbol{y_o}+\boldsymbol{z_i}\|^2 \\
&= \frac{1}{p}\sum_{i=1}^{p}\left[\frac{1}{l}\sum_{j=1}^{l}(\overline{y_j}-y_{oj}+z_{ij})^2\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\left[\frac{1}{l}\sum_{j=1}^{l}(\overline{y_j}-y_{oj})^2+z_{ij}^2+2z_{ij}(\overline{y_j}-y_{oi})\right] \\
&= \frac{1}{p}\sum_{i=1}^{p}\left[\|\bar{\boldsymbol{y}}-\boldsymbol{y_o}\|^2+\frac{1}{l}\sum_{j=1}^{l}z_{ij}^2+\frac{2}{l}\sum_{j=1}^{l}z_{ij}(\overline{y_j}-y_{oj})\right] \\
&= \|\bar{\boldsymbol{y}}-\boldsymbol{y_o}\|^2+D^2+\frac{1}{p}\sum_{i=1}^{p}\frac{2}{l}\sum_{j=1}^{l}(y_{ij}-\overline{y_j})(\overline{y_i}-y_{oj}) \\
&= \|\bar{\boldsymbol{y}}-\boldsymbol{y_o}\|^2+D^2+\frac{1}{p}\sum_{i=1}^{p}\frac{2}{l}\sum_{j=1}^{l}(y_{ij}\overline{y_j}-\overline{y_j^2}-y_{ij}y_{oj}+\overline{y_j}y_{oj}) \\
&= \|\bar{\boldsymbol{y}}-\boldsymbol{y_o}\|^2+D^2+\frac{2}{l}\sum_{j=1}^{l}\frac{1}{p}\sum_{i=1}^{p}(y_{ij}\overline{y_j}-\overline{y_j^2}-y_{ij}y_{oj}+\overline{y_j}y_{oj}) \\
&= \|\bar{\boldsymbol{y}}-\boldsymbol{y_o}\|^2+D^2 \qquad \text{(식 3.36)}
\end{aligned}
$$

식 3.35에 나타나 있듯이 전체의 오차는 앙상블이 퍼져있는 정도와 앙상블 평균시나리오가 가지는 오차로 나누어진다. 그러므로 각각의 앙상블이 가지는 전체 오차를 나타내는 MSE, S^2, 은 항상 평균시나리오의 오차보다 커지게 된다. 이것이 앙상블 평균 예측을 사용하는 이유가 된다(Epstein, 1969; Leiith, 1974; Brankovic et al., 1990; Stephenson and Doblas-Reyes, 2000).

3.2.2 다중범주 확률예측 평가방법

앙상블 예측의 정확도를 평가하기 위하여 확률예측 평가방법으로 BS(Brier Score)가 있다. BS는 본질적으로 확률예측의 평균제곱오차(Mean Squared Error, MSE)를 의미한다(Wilks, 2006). 일반적으로 사용하는 BS는 다음 식 3.37과 같다.

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(f_k - o_k)^2 \tag{3.37}$$

여기서 f_k는 k 번째의 예측의 확률 값을 의미하며, o_k는 k 번째 관측의 확률 값을 의미한다. 여기서 시계열에서 k는 시간으로 간주하여도 무방하다. o_k는 관측의 확률 값이므로 사건이 일어났을 경우 1로, 사건이 일어나지 않았을 경우에는 0의 값을 가지게 된다. 예를 들어 50 mm 이상의 강수가 내릴 확률을 70%라고 예측을 하였고, 실제로 50 mm 이상의 강수가 내렸다면 BS는 $(0.7-1)^2$이 되어 0.09가 된다. 50 mm 이하에 대한 평가는 하지 않게 되므로 BS는 0부터 1까지의 범위를 가지게 되며, 정확도가 높은 예측을 할수록 0에 가까운 값을 가지게 된다.

위에서 언급한 BS는 이항 사건(binary event)에 대한 예측에 사용이 적합한 식이다. 본 연구에서는 Brier(1950)가 제안한 본래의 BS를 사용하였다. 그 식은 다음과 같다(식 3.38).

$$\mathrm{BS} = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{r}(f_{ti} - o_{ti})^2 \tag{3.38}$$

여기서 f_{ti}는 예측의 확률 값을 나타내며 o_{ti}는 관측의 확률 값을 나타낸다. n은 식 3.37의 n과 같아 시간으로 볼 수 있다. 그리고 r은

예측구간을 나타낸다. 식 3.37의 예시에서는 50 mm이상의 구간만을 예측하였으므로 r이 1인 경우이다. 식 3.38에서는 r이 1이 아닌 값을 가지게 되어 다중범주에 관한 평가가 가능하게 된다. 또한 이 경우에는 BS가 0부터 2까지의 범위를 가지게 된다. 이를 구분하기 위하여 식 3.37의 BS를 half-Brier score 라고 부른다(Wilks, 2006).

본 연구는 어떤 특정 구간에 대한 예측을 평가하는 것이 아니라 시나리오 간 상관성이 앙상블 시스템에 미치는 영향을 알아보기 위한 것이므로 식 3.38의 BS를 사용하였다. 또한 BS는 r을 어떻게 설정하느냐에 따라 예측의 정확도가 다르게 나타날 수 있다. 본 연구에서는 시나리오 수가 3개를 사용한 것이 최소 사용 시나리오 수이므로 r 설정의 영향을 줄이기 위하여 r을 3으로 설정하였다. 또한 표준정규분포에서 난수를 발생하여 모의를 하였으므로 표준정규분포를 동일한 확률로 삼등분하여 구간의 기준을 정하였다.

3.2.3 유효 앙상블 시나리오 수 추정방법

앙상블 시나리오 수와 앙상블 예측 시스템과의 관계를 알아보고 적절한 수의 시나리오 수를 제시하기 위하여 2가지 방법을 사용하였다. BS를 사용하여 앙상블 예측 시스템이 효과적으로 향상되는 앙상블 멤버 수를 찾으려고 하였다. 유효 앙상블 시나리오 수는 적절한 예측의 정확도를 가지는 최소한의 시나리오 수로 정의하였다. 많은 수의 시나리오를 사용할수록 예측의 정확도가 향상되는 경우에 앙상블 모의 시간과 데이터의 양을 효율적으로 조절하기 위하여 적절한 수의 시나리오가 제시될 필요가 있다.

BS가 시나리오 수의 증가에 따라 향상되는 정도를 구간별로 비교하여 분석하였다. 첫 번째 방법은 스코어의 최대값과 최소값의 차이의 90% 되는 지점을 추정하여 개수를 정하였다. 예를 들어 앙상블 시나리오 수가 증가할수록 예측이 향상되는 경우, 3개의 시나리오를 사용한 앙상블과 100개를 사용한 앙상블 스코어의 차이의 90%를 계산하였다. 그리고 그 스코어에 해당하는 앙상블 수를 추정하였다.

두 번째 방법은 BS가 향상되는 정도를 그래프의 기울기로 판단하는 방법이다. 일정이상의 기울기, 즉 향상 정도를 가지지 못하는 지점에서 적절한 수의 시나리오라고 판단하였다. 대부분의 경우에서 3개에서 5개로 시나리오 수가 늘어났을 때 가장 스코어가 크게 향상되었다. 이때의 기울기의 5%가 되는 지점을 앙상블 멤버 수 판단의 근거로 사용하였다.

제 4 장 적용 및 결과

4.1 개요

시나리오 간 상관성(ρ_e)이 앙상블 예측 시스템에 미치는 영향을 분석하기 위하여 6 종류의 ρ_e를 이용하여 상관 앙상블을 생성하여 예측 능력을 평가하였다. 또한 상관성에 따른 유효 앙상블 시나리오 수를 찾기 위하여 ρ_e별로 3부터 100까지 10종류의 시나리오 수를 이용하였다. 상관 앙상블을 첫 번째 시나리오와 관측시나리오의 상관계수인 ρ_a는 다양한 상황에 대하여 모의하기 위하여 사용하였다.

4.2 변동계수 효과 분석에서는 표준정규분포 사용의 타당성을 확인하기 위하여 변동계수(C_v)를 0.1, 0.5, 1 세 가지로 조절하여 상관 앙상블을 생성하였다. 4.2절에서는 각 종류의 ρ_a와 C_v별로 앙상블을 SE 지표로 평가하였다.

4.3 앙상블 시나리오 간 상관성 효과 분석에서는 본격적으로 앙상블 시나리오를 평가하였다. 평가를 위하여 확정적 예측 평가기법과 확률예측 평가기법으로 나누어 평가를 하였다. 확정적 예측 평가기법으로 평균 앙상블의 평균제곱오차인 SE와 앙상블 시나리오가 퍼진 정도를 나타내는 D^2를 사용하였다. 확률예측 평가를 위하여 BS를 사용하였다.

각 평가기법별로 다시 전체 결과 분석과 ρ_a별 결과 분석으로 나뉜다. 전체 결과 분석은 ρ_a별로 나누지 않고 ρ_e만으로 분리해서 평가한 것이다. 그리고 ρ_a별 결과 분석은 ρ_a별로 나누어 상관 앙상블을 평가한 것이다. ρ_a를 고려하지 않은 분석은 앙상블을 사용할 시 예측의 정확도를 예상할 수 없는 경우를 나타낸 것이다. 이 경우는 시나리오의 상관성만을 고려하여 결과를 분석하였다. ρ_a를 고려한

분석은 정확한 예측을 수행할 수 있는 경우와 불확실성이 아주 높아 비교적 부정확한 예측을 수행할 수밖에 없는 상황을 모의한 것이다. ρ_a가 크다면 전자에 해당되며 ρ_a가 낮은 경우가 후자에 해당된다.

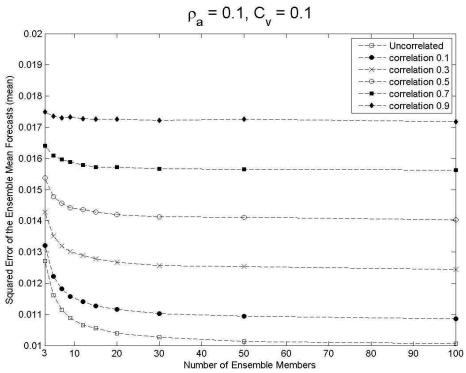
4.4 유효 앙상블 시나리오 수 추정에서는 유효 앙상블 시나리오 수를 제시하기 위하여 4.3 절의 BS 결과를 이용하였다. BS 지표가 향상되는 정도를 2 가지 방법을 이용하여 분석하였다. 첫 번째는 BS의 최대값과 최소값의 차이의 비율을 구하는 방식이며 두 번째 방법은 시나리오가 증가 대비 지표의 향상 정도를 분석하였다. 최적의 예측 정확도를 보이는 시나리오 수가 아닌 적절한 정확도를 보이는 최소한의 시나리오 수를 찾는 것에 목적을 두었다.

4.2 변동계수 효과 분석

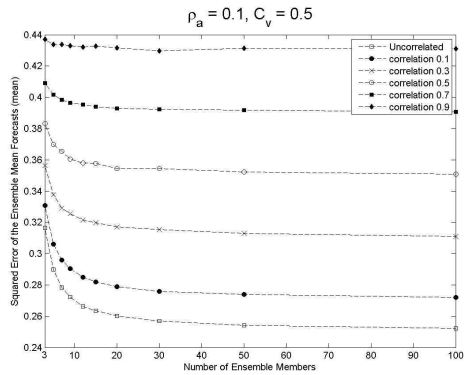
표준정규분포의 사용 가능 여부를 판단하기 위하여 McMahon et al.(2007)이 제시한 세계 1221개 강들의 연홍수량의 변동계수를 이용하였다. 실제 유량에 맞는 변동계수를 이용하여 본 연구에서 사용한 분포의 적절성을 확인하고자 한다. 변동계수 C_v 값은 제시된 범위 내에서 결정하여 0.1, 0.5, 그리고 1로 정하였다. 그리고 앙상블 예측 시스템의 정확도(ρ_a)를 다르게 설정하여 각 상황별로 변동계수의 영향을 분석하였다.

그림 4.1은 앙상블 예측 시스템의 정확도가 0.1인 경우의 결과이다. x축은 앙상블 시나리오 수를 나타내며 y축은 SE를 나타낸다. 그래프에 나타난 값은 SE를 평균하여 나타내었다. 그림 4.1(a)는 변동계수가 0.1, 그림 4.1(b)는 변동계수가 0.5, 그림 4.1(c)는 변동계수가 1일 경우의 결과를 보여준다. 세 가지 그림 모두 앙상블 시나리오 수가 증가할 때 SE가 작아지는 것을 확인할 수 있다. 또한 앙상블 시나리오 간 상관계수가 작아질수록 SE가 작아진다. 그림 4.2는 앙상블 예측 시스템의 정확도가 0.3인 경우의 결과이다. 정확도가 0.1인 경우와 같은 결과가 나타났다. 변동계수가 달라져도 그래프의 개형이 변하지 않는 것을 확인할 수 있다.

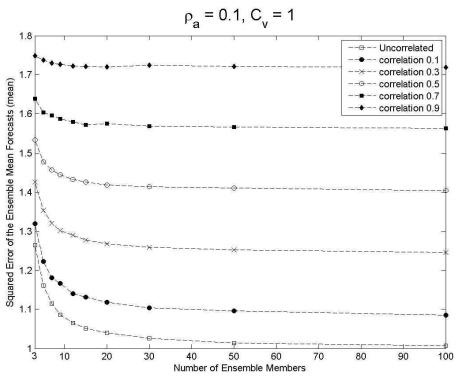
그림 4.3은 앙상블 예측 시스템의 정확도가 0.5인 경우의 결과이다. 변동계수가 달라져도 앙상블 시나리오 수와 시나리오 간 상관계수가 SE에 영향을 미치지 못하고 무질서한 양상을 보여준다. 그림 4.4와 4.5는 시스템의 정확도가 각각 0.7과 0.9인 경우의 결과이다. 세 가지 변동계수 결과 모두 시나리오 수가 증가할수록, 시나리오 간 상관계수가 작아질수록 SE가 커지는 것을 확인할 수 있다. 변동계수에 변화가 있어도 결과 그래프 개형에 변화가 없고 지표의 크기만 다르게 되는 것으로 나타났다. 따라서 표준정규분포를 사용하는 것에 문제가 없음을 확인하였다.



(a)

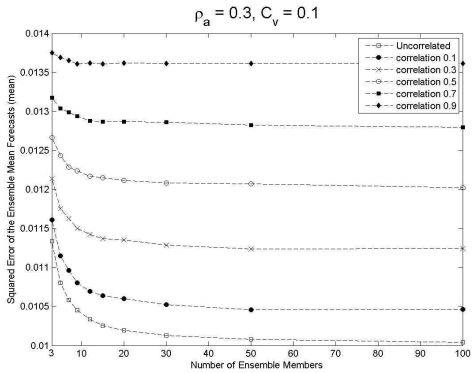


(b)

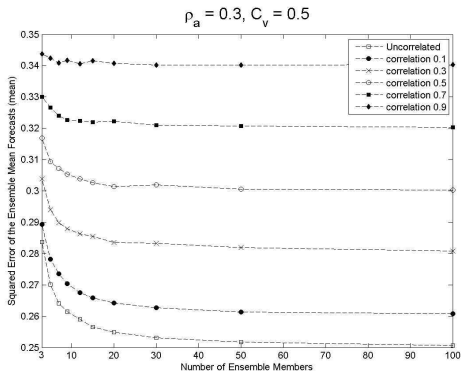


(c)

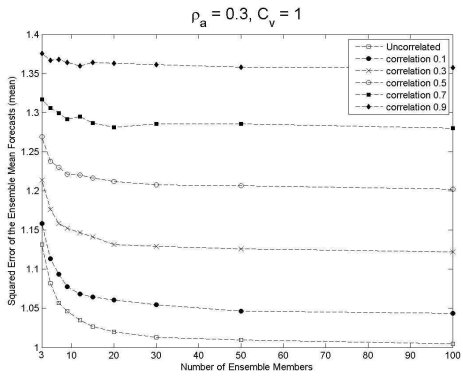
Fig. 4.1 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.1$



(a)

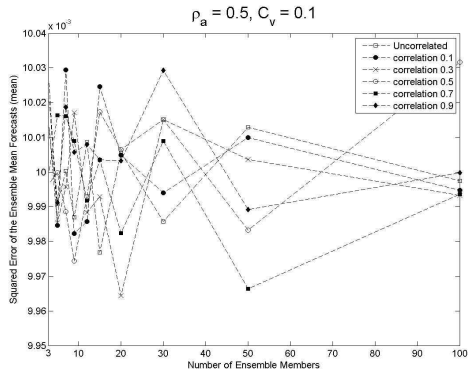


(b)

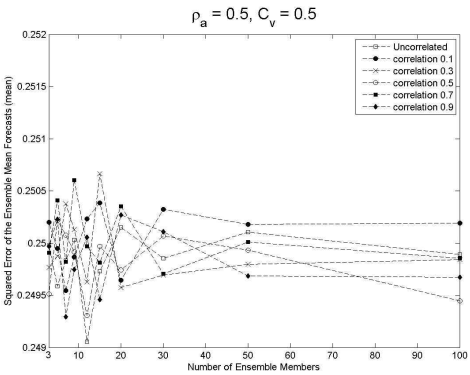


(c)

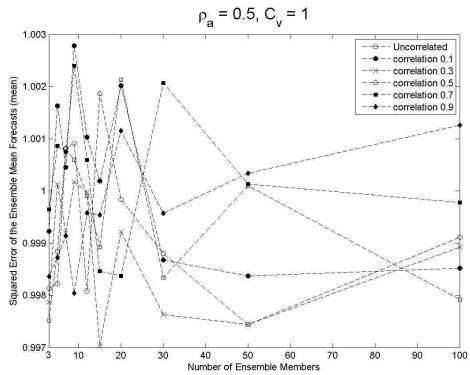
Fig. 4.2 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.3$



(a)

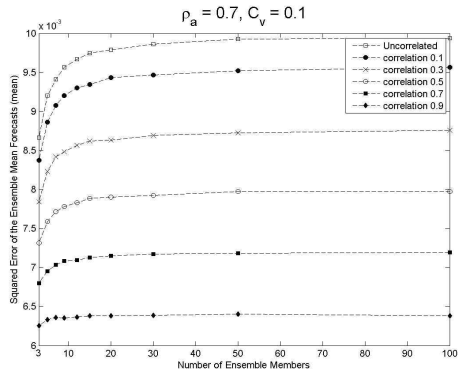


(b)

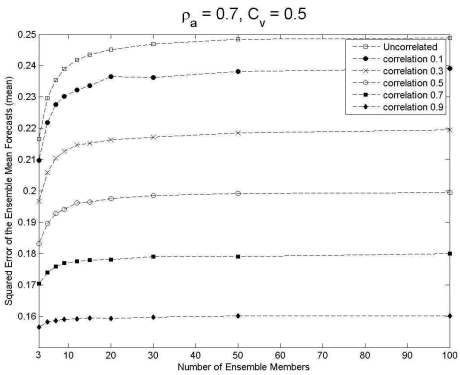


(c)

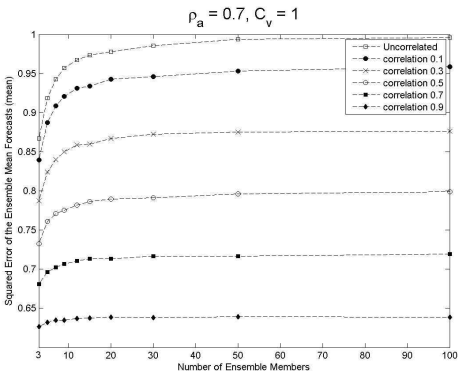
Fig. 4.3 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.5$



(a)

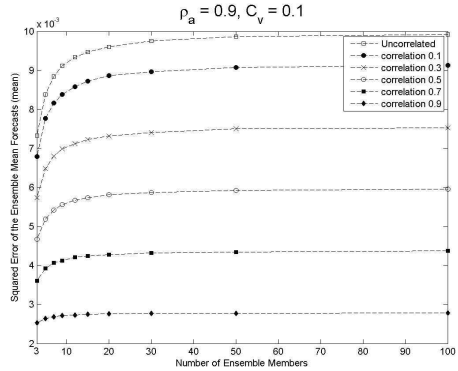


(b)

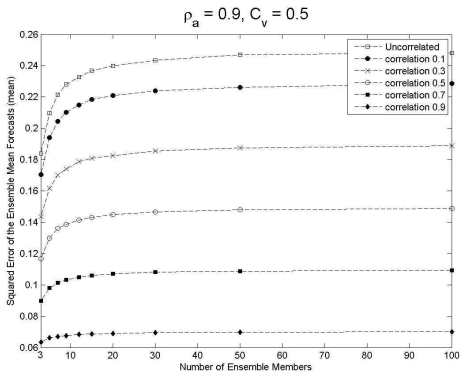


(c)

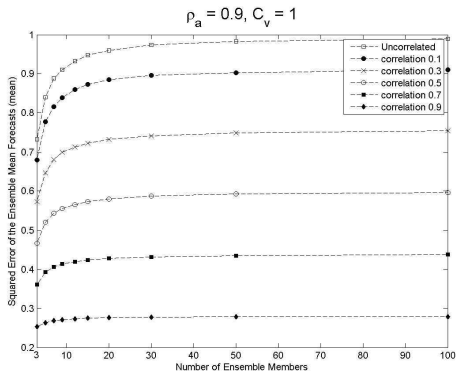
Fig. 4.4 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.7$



(a)



(b)



(c)

Fig. 4.5 Squared error of the ensemble mean forecasts(mean), $\rho_a = 0.9$

4.3 앙상블 시나리오 간 상관성 효과분석

4.3.1 확정적 예측 평가

(1) 전체 결과 분석

앙상블 시나리오 간 상관성(ρ_e)의 효과를 분석하기 위하여 확정적 예측 평가 방법으로써 평가지표 SE와 D^2를 사용하였다. SE는 평균 앙상블 시나리오와 관측시나리오의 오차를 나타내는 지표이며, D^2는 앙상블 시나리오들이 평균 시나리오로부터 얼마나 퍼져있는가를 평가하는 지표이다. 두 지표의 합인 MSE는 각 시점에서의 확률예측이 아닌 각 시나리오와 관측시나리오의 오차를 제곱한 것의 합이므로 본 연구에서 지표로 사용하는 것으로 부적절하다고 판단하였다. 그리고 D^2는 관측시나리오와 앙상블 시나리오 간의 관계에 무관하게 앙상블 예측 시스템 자체에 관한 특성을 평가하는 지표이므로 ρ_a별로 분석을 하는 경우에도 앙상블 시나리오 간 상관계수가 같다면 D^2가 같게 나타난다. 따라서 D^2 결과는 전체 결과 분석 절에서만 서술하였다.

그림 4.6부터 그림 4.11은 시나리오 간 상관성이 앙상블 예측 시스템에 미치는 영향을 SE를 사용하여 분석한 결과이다. 상관성의 영향만을 확인하기 위하여 ρ_a별 결과 중 같은 시나리오 간 상관계수를 가지는 결과를 이용하여 각각의 그래프를 그렸다. x축은 앙상블 시나리오 수를 나타내며 y축은 SE를 나타낸다. 그림 4.6은 시나리오 간 상관계수가 0인 경우를 나타낸다. 시나리오 수가 증가할수록 boxplot의 크기가 작아지는 것을 확인할 수 있다. 즉 SE 결과의 분산이 시나리오 수가 증가할수록 작아진다. 그런데 결과의 중앙값이 시나리오 수에 관계없이 거의 1 값에 고정되어있다. 따라서 평균 앙

상블 시나리오를 사용했을 시에 시나리오 수가 많을수록 안정적으로 더 좋은 예측을 할 수 있다는 것을 알 수 있다. 시나리오 수가 적을 때는 많은 경우보다 더 좋은 예측을 할 수도 있지만 오차가 큰 예측을 수행할 수도 있기 때문에 시나리오 수가 적을수록 불확실성이 크다고 할 수 있다. 또한 시나리오 수가 20개부터는 수가 증가해도 boxplot의 큰 변화가 나타나지 않는다. 시나리오 간 상관계수가 0이고 평균 앙상블 시나리오를 사용할 경우 멤버 수가 20개 이후부터는 멤버 수 증가가 의미 없음을 보여준다.

그림 4.7은 시나리오 간 상관계수가 0.1인 경우의 결과이다. 상관계수가 0인 경우에 비하여 boxplot의 크기 더 커진 것을 확인할 수 있다. 따라서 시나리오 간 상관계수의 증가로 앙상블 예측 시스템이 가질 수 있는 정보의 손실로 인해 불확실성이 커진 것을 보여준다. 그리고 시나리오 간 상관성이 커질수록 이와 같은 경향성이 증가하는 것을 알 수 있다. 시나리오 간 상관성이 커질수록 불확실성이 커지며 시나리오 수 증가의 효과가 점점 약해지는 것이 나타난다. 특히 시나리오 간 상관계수가 0.9인 경우(그림 4.11), 시나리오 수 증가가 앙상블 예측 정확도에 아무런 영향을 미치지 못하는 것을 확인할 수 있다.

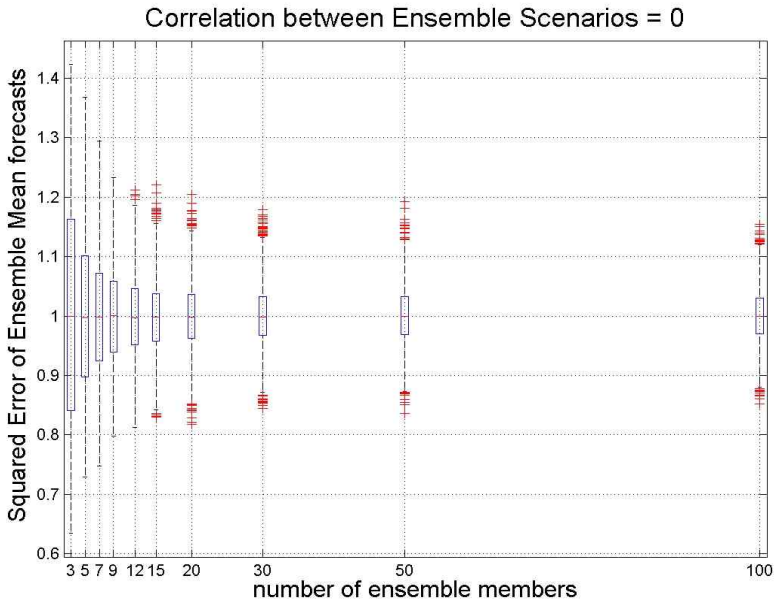


Fig 4.6 Squared error of ensemble mean forecasts (ρ_e =0)

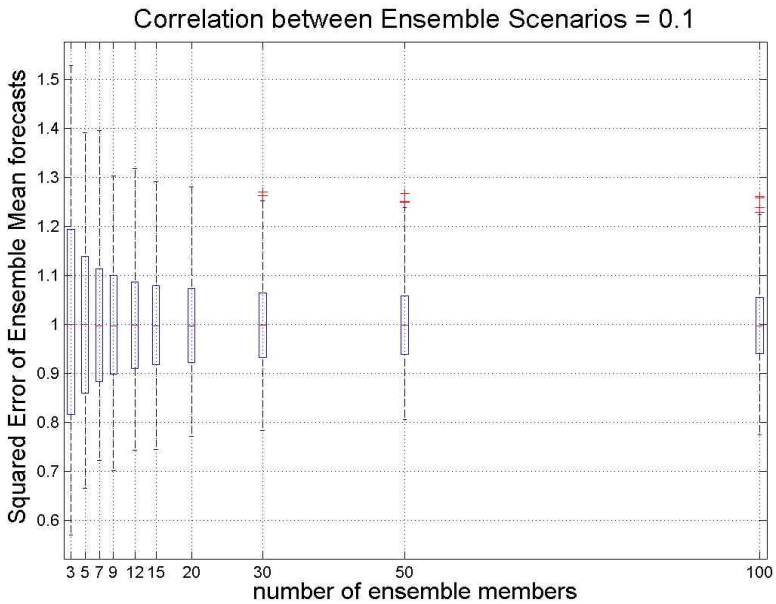


Fig 4.7 Squared error of ensemble mean forecasts (ρ_e =0.1)

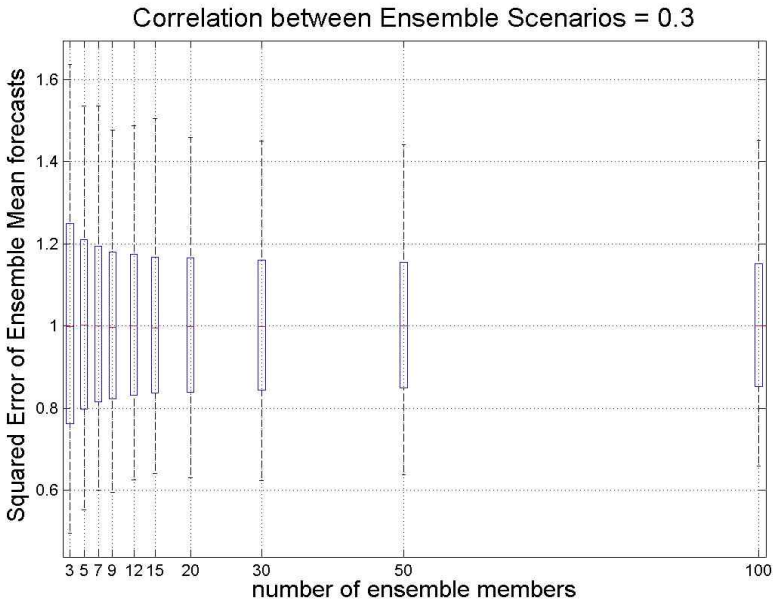


Fig. 4.8 Squared error of ensemble mean forecasts (ρ_e =0.3)

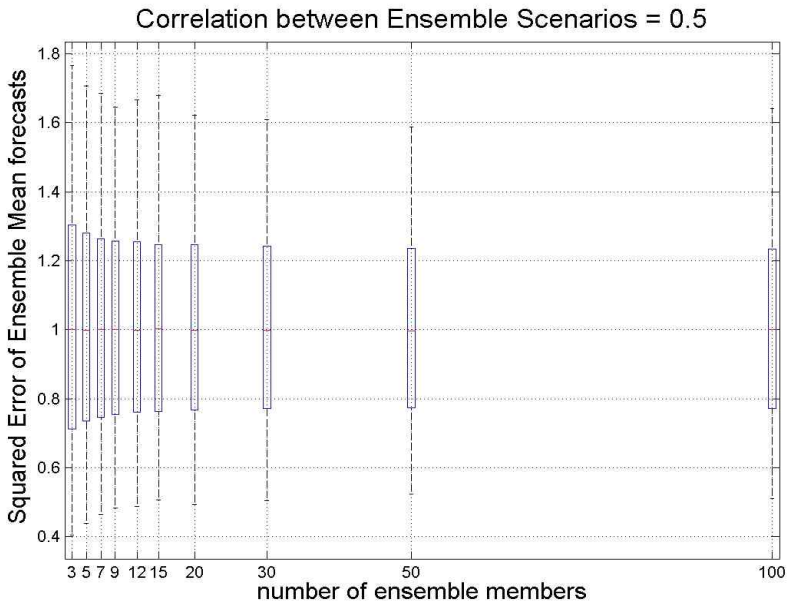


Fig. 4.9 Squared error of ensemble mean forecasts (ρ_e =0.5)

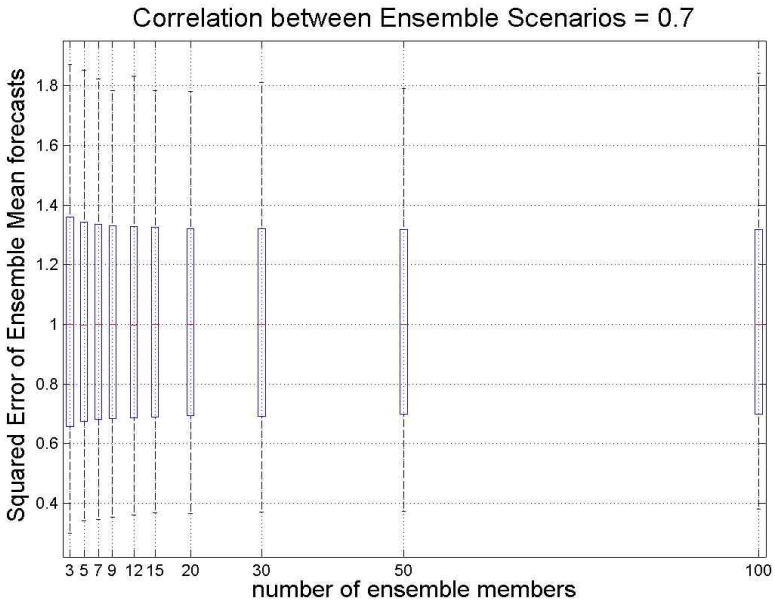


Fig. 4.10 Squared error of ensemble mean forecasts (ρ_e =0.7)

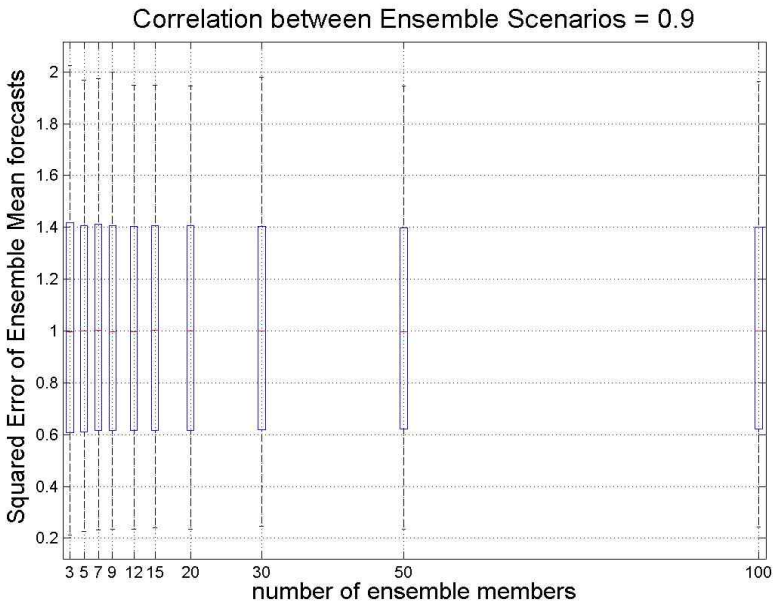


Fig. 4.11 Squared error of ensemble mean forecasts (ρ_e =0.9)

그림 4.12부터 그림 4.28은 mean squared dispersion(D^2)를 사용하여 앙상블 시나리오 간 상관성(ρ_e)과 앙상블 시나리오 수가 앙상블 시나리오 간 거리에 미치는 영향을 분석한 결과이다. 식 3.35에 나타나 있듯이 전체 시스템의 오차는 SE와 D^2의 합으로 이루어져 있다. 그러므로 앙상블 예측에서 D^2가 증가하면 평균 앙상블 예측의 오차가 감소할 수 있음을 의미한다. 일반적으로 앙상블에서 D^2가 클수록 좋은 예측 시스템이라고 할 수 있다. 그림 4.13부터 그림 4.22의 x축은 ρ_e를 나타내며 y축은 D^2 결과이다. 그리고 결과를 앙상블 시나리오 수별로 나타내었다. ρ_e가 0에서 0.9까지 커짐에 따라서 D^2가 작아지는 것을 확인할 수 있다. 상관성이 커질수록 앙상블 시나리오 간 거리가 가까운 것을 보여준다. 또한 ρ_e가 커질수록 boxplot의 크기가 작아지고 있으며, 이것은 시나리오 간 상관성이 커질수록 가능한 D^2값의 범위가 작아진다는 것을 의미한다. 따라서 상관성의 증가는 표현할 수 있는 예측 범위를 감소시킨다.

그림 4.23부터 그림 4.28에서 x축은 시나리오 수를 나타내고 y축은 D^2를 나타낸다. 그리고 결과를 앙상블 시나리오 간 상관성별로 도시하였다. 같은 시나리오 간 상관성을 가진 경우, 시나리오 수가 증가할수록 D^2가 커지는 것을 볼 수 있다. 이 결과를 평균하여 정리하면 그림 4.12와 같다. 이 그래프를 살펴보면 시나리오 수가 적은 구간에서, 시나리오 간 상관성이 작을수록 그래프의 기울기가 크다는 것을 알 수 있다. 시나리오 수가 작은 구간에서 수가 증가할 때 D^2가 증가하는 정도가 커지며, 일정 수 이상에서는 시나리오 수 증가가 앙상블 시스템이 표현할 수 있는 예측 범위 증가에 크게 영향을 미치지 못함을 보여준다.

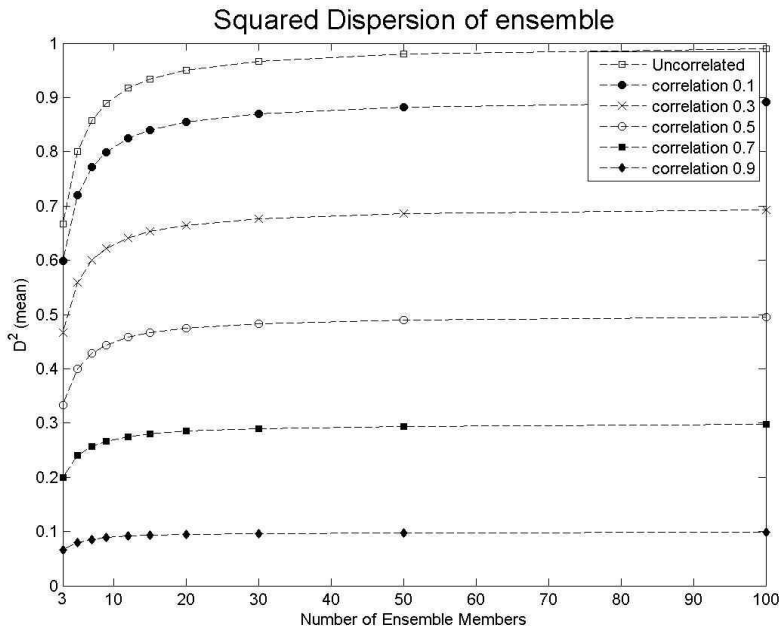


Fig. 4.12 Mean squared dispersion (summary)

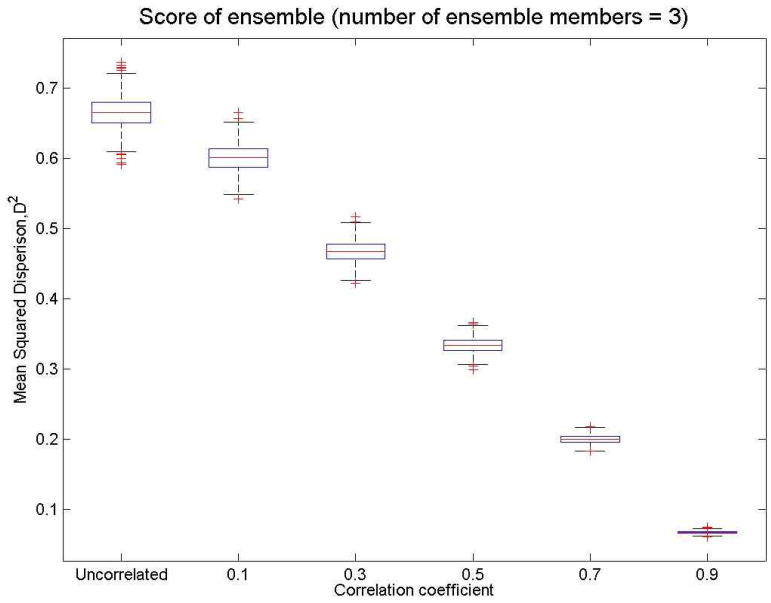


Fig. 4.13 Mean squared dispersion(ensemble members, 3)

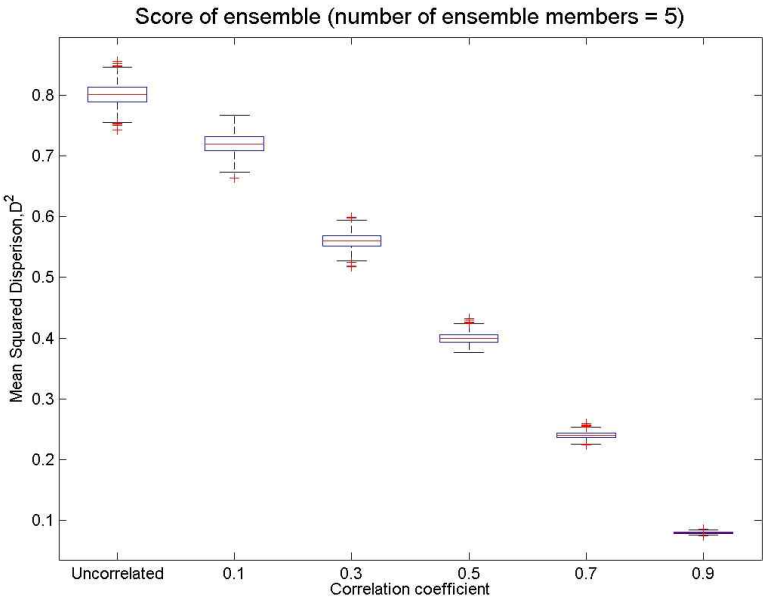


Fig. 4.14 Mean squared dispersion(ensemble members, 5)

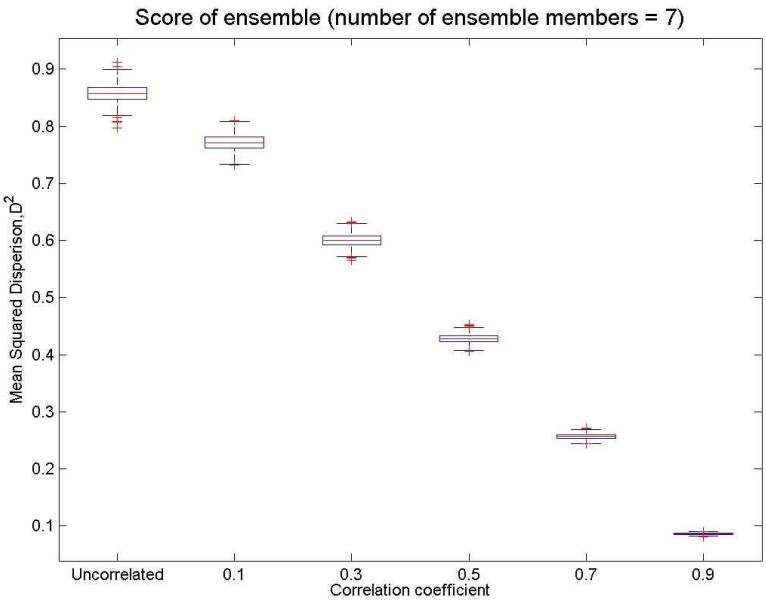


Fig. 4.15 Mean squared dispersion(ensemble members, 7)

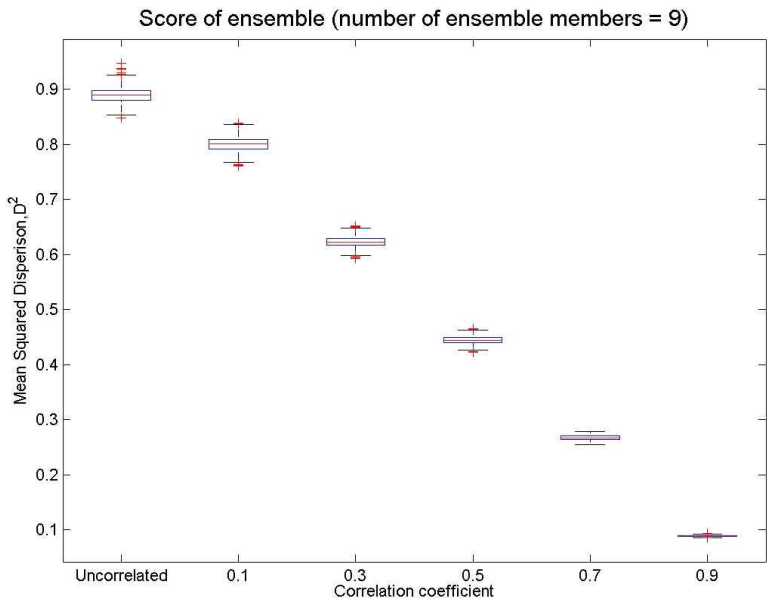


Fig. 4.16 Mean squared dispersion(ensemble members, 9)

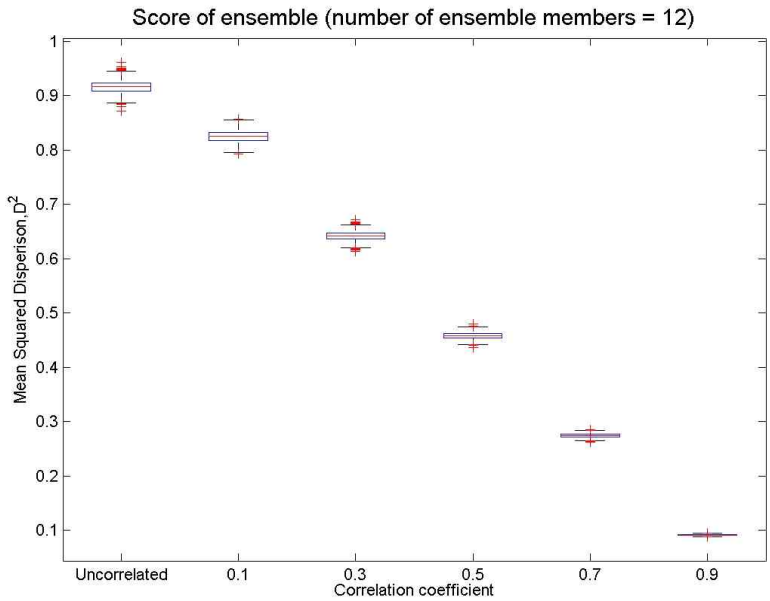


Fig. 4.17 Mean squared dispersion(ensemble members, 12)

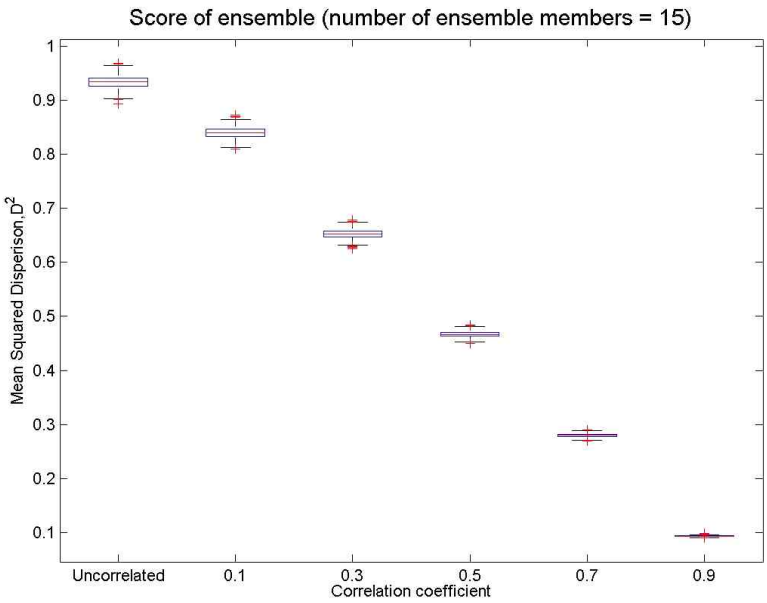


Fig. 4.18 Mean squared dispersion(ensemble members, 15)

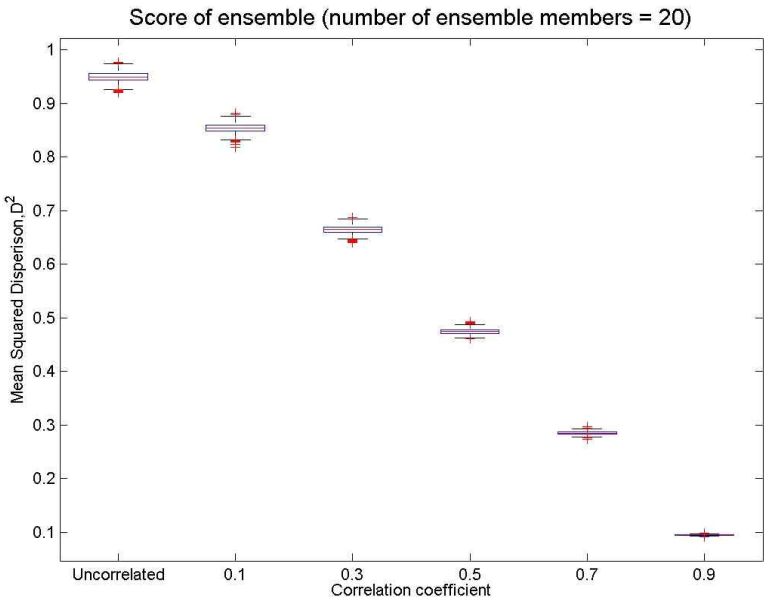


Fig. 4.19 Mean squared dispersion(ensemble members, 20)

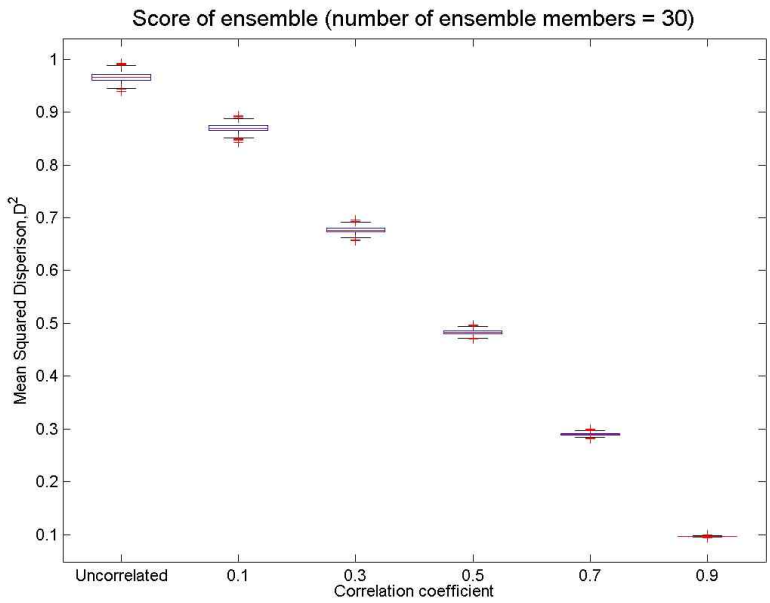


Fig. 4.20 Mean squared dispersion(ensemble members, 30)

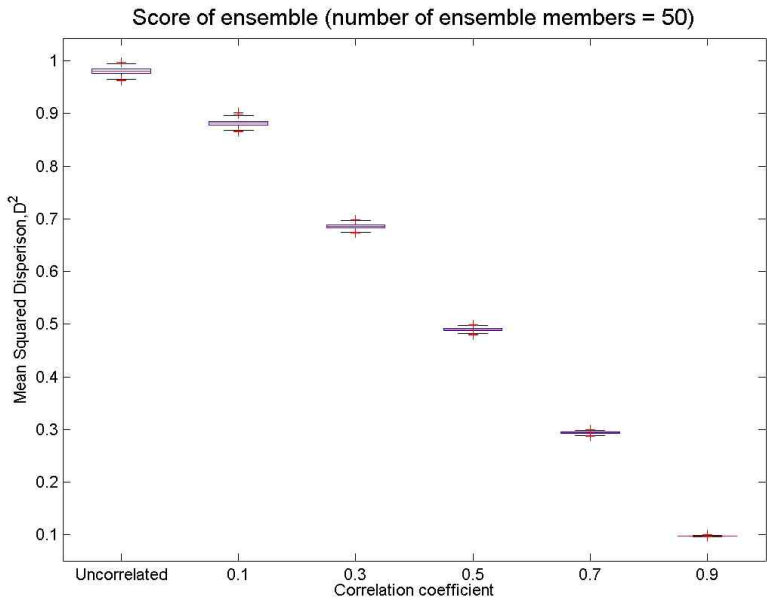


Fig. 4.21 Mean squared dispersion(ensemble members, 50)

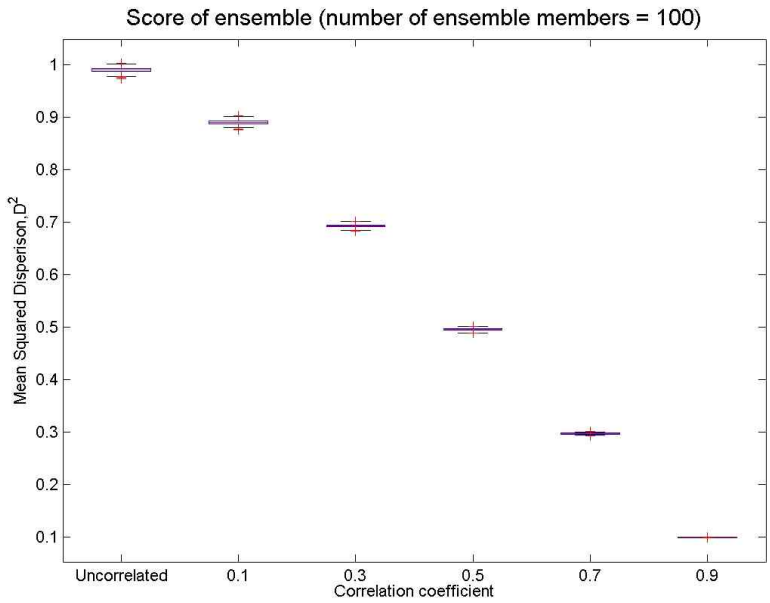


Fig. 4.22 Mean squared dispersion(ensemble members, 100)

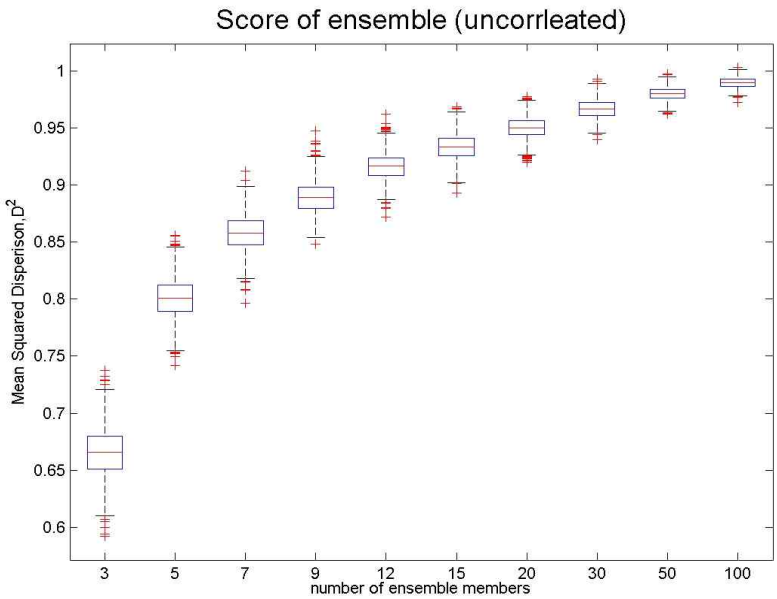


Fig. 4.23 Mean squared dispersion(uncorrelated, $\rho_e = 0$)

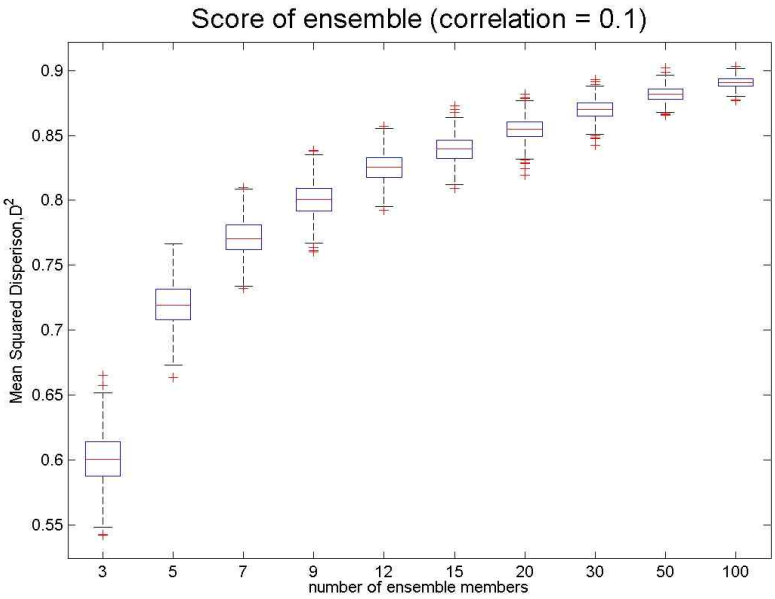


Fig. 4.24 Mean squared dispersion(ρ_e =0.1)

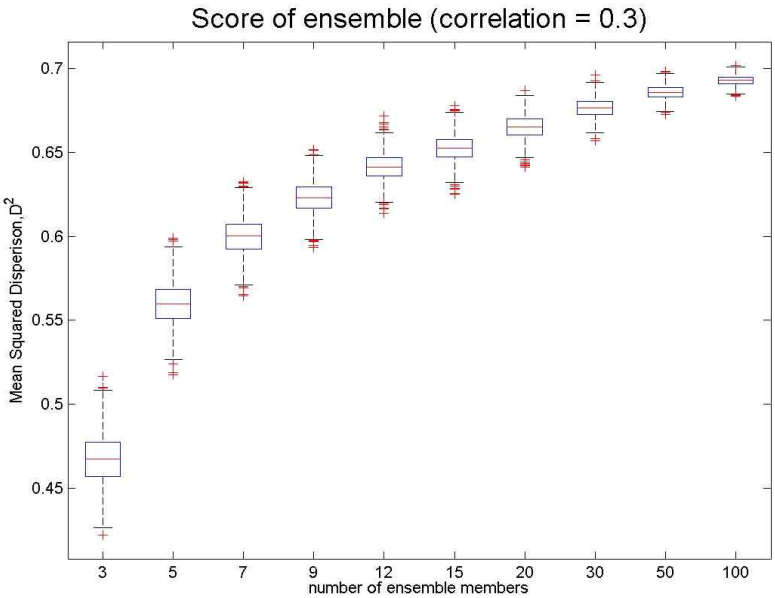


Fig. 4.25 Mean squared dispersion(ρ_e =0.3)

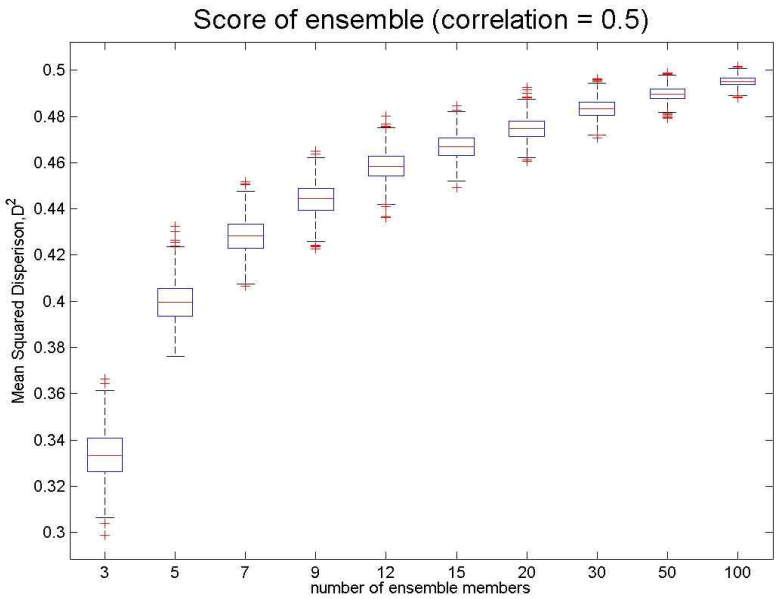


Fig. 4.26 Mean squared dispersion(ρ_e =0.5)

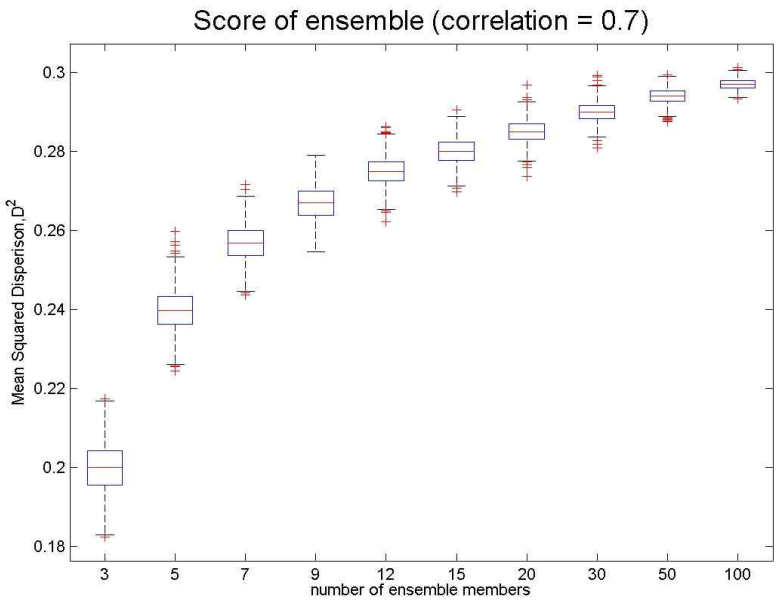


Fig. 4.27 Mean squared dispersion(ρ_e =0.7)

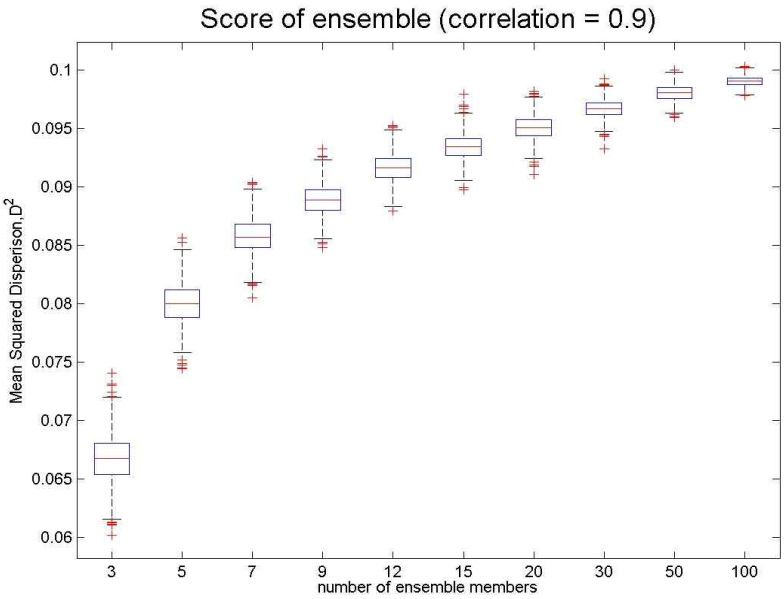


Fig. 4.28 Mean squared dispersion(ρ_e =0.9)

(2) ρ_a별 결과 분석

본 절에서는 앙상블 예측 평가를 위하여 생성한 관측시나리오의 생성 시에 사용한 상관계수(ρ_a, nominal accuracy)별로 결과를 분석하였다. ρ_a는 앞서 언급한 바와 같이 상관 앙상블 행렬 $\boldsymbol{Y}$의 첫 번째 열이며 시나리오인 $\boldsymbol{y_1}$와 관측시나리오 $\boldsymbol{y_o}$의 상관계수이다.

ρ_a는 앙상블의 한 시나리오와 관측시나리오 사이의 상관계수, 즉 정확도를 나타내는 지표이기 때문에 앙상블 전체의 정확도를 표현하는데 한계가 있을 수 있다. 하지만 시나리오 $\boldsymbol{y_1}$는 상관 앙상블 생성 시에 모든 시나리오에 영향을 끼치며, 초기의 앙상블 데이터 행렬 $\boldsymbol{X}$에서 변하지 않는 유일한 시나리오이다. 그러므로 시나리오 $\boldsymbol{y_1}$에 대표성이 있다고 판단할 수 있다. 또한 $\boldsymbol{y_1}$과 상관 앙상블의 나머지 시나리오들과의 관계를 알고 있으며, 각 경우별로 1000회씩 모의를 하고 있다. 따라서 ρ_e와 ρ_a는 앙상블 예측 시스템 내에 비교적 정확한 예측 시나리오가 존재하는지에 대한 유무와 정확한 예측 시나리오와 얼마나 유사한 퍼포먼스를 보이는 시나리오가 존재하는지에 대한 판단을 할 수 있게 한다.

① $\rho_a = 0.1$

그림 4.29부터 4.34는 ρ_a가 0.1인 경우의 SE 결과를 boxplot으로 나타낸 것이다. x축은 앙상블 시나리오 수를 나타내며 y축은 평균 앙상블 시나리오의 오차를 나타내는 SE 결과이다. 그리고 시나리오 간 상관성(ρ_e)별로 결과를 나타내었다. 그림 4.29는 ρ_a가 0.1이며 ρ_e가 0인 경우이다. 이 경우는 부정확한 예측 시나리오가 앙상블 내에 존재하며 앙상블 시나리오 간 상관성이 없는 상황을 나타낸다. 앙상블 시나리오 수가 증가할수록 오차가 작아지는 것을 확인할 수 있

다.

ρ_e가 증가하면 전반적인 오차의 크기가 커지는 것을 확인할 수 있다(그림 4.30, 4.31, 4.32, 4.33, 그리고 4.34). 비교적 부정확한 예측을 수행하는 경우 앙상블 간 상관성이 증가하면 오차가 전체적으로 증가하는 경향을 볼 수 있다. 또한 앙상블 간 상관성이 증가하면 시나리오 수 증가에 따라 오차가 줄어드는 정도가 작아지는 것을 볼 수 있다. 앙상블 간 상관성이 아주 높은 경우(ρ_e= 0.9), 시나리오 수 증가의 효과가 거의 확인되지 않는다. 상관성 증가로 인한 정보의 손실 효과가 시나리오 수 증가 효과에 비해 더 크게 작용하고 있는 것으로 보인다.

또한 같은 수의 앙상블 시나리오 수를 가지는 경우에 ρ_e가 증가할수록 box의 크기가 커지는 것을 볼 수 있다. 시나리오 간 상관성이 커질수록 평균 앙상블 예측 오차의 분산이 커지는 것을 알 수 있다. 그리고 오차의 평균이 상관성 증가와 함께 커지므로 상관성이 커질수록 더 큰 오차를 발생시킬 가능성이 높아질 것으로 보인다.

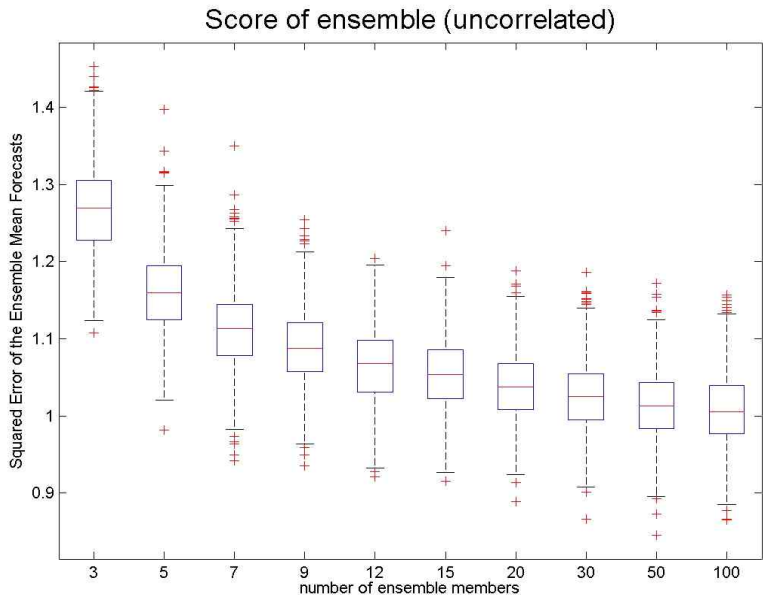


Fig. 4.29 SE of the ensemble mean forecasts (ρ_e =0, ρ_a = 0.1)

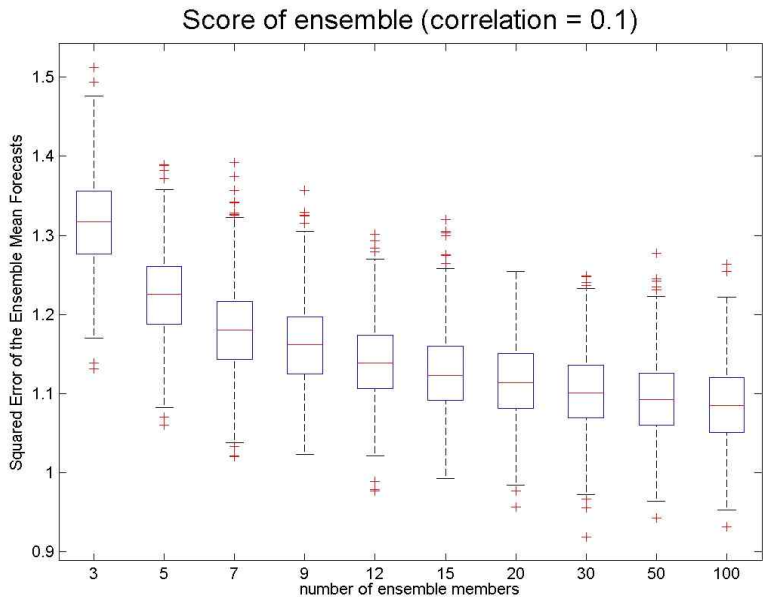


Fig. 4.30 SE of the ensemble mean forecasts (ρ_e =0.1, ρ_a = 0.1)

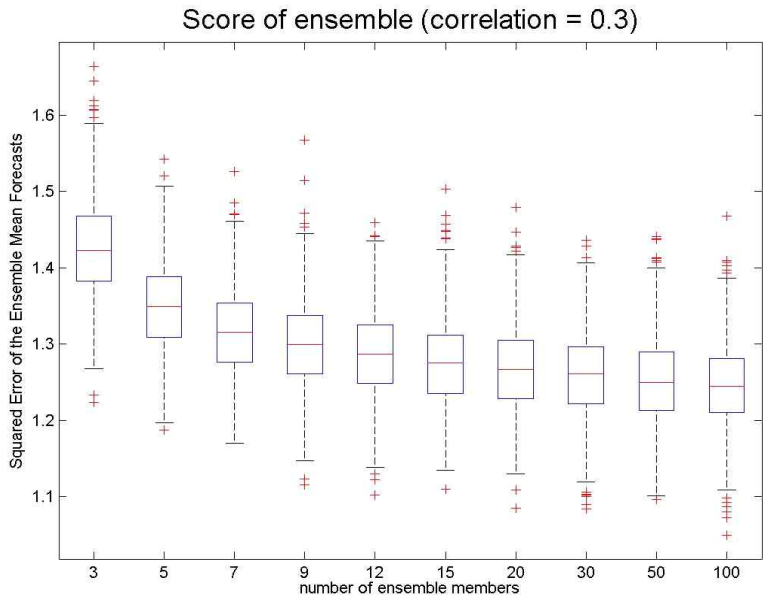


Fig. 4.31 SE of the ensemble mean forecasts (ρ_e =0.3, ρ_a = 0.1)

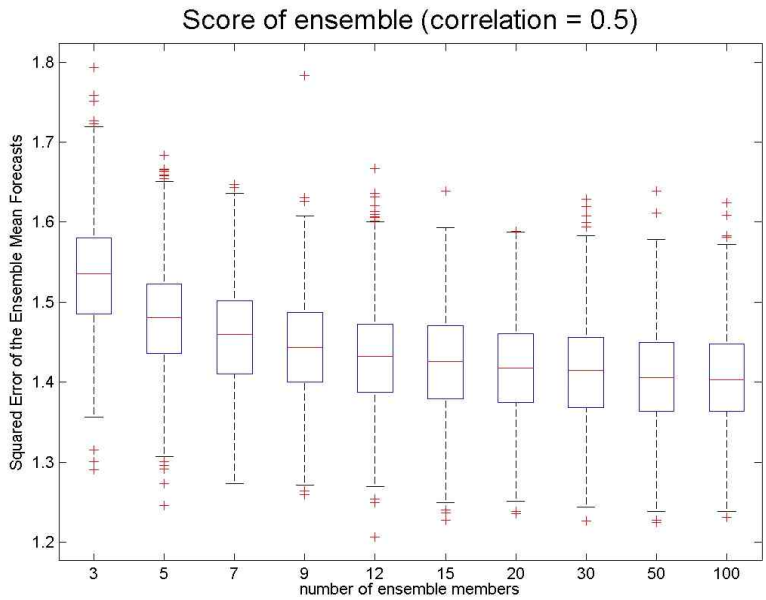


Fig. 4.32 SE of the ensemble mean forecasts (ρ_e =0.5, ρ_a = 0.1)

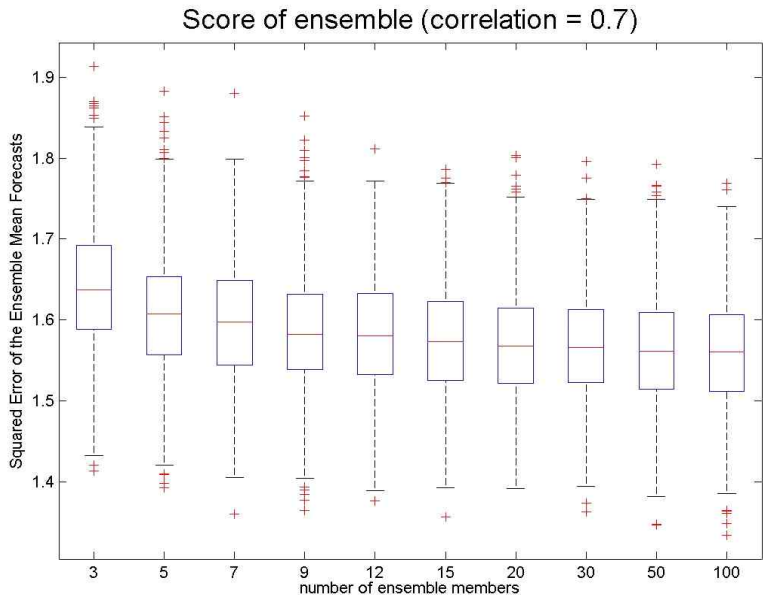


Fig. 4.33 SE of the ensemble mean forecasts (ρ_e =0.7, ρ_a = 0.1)

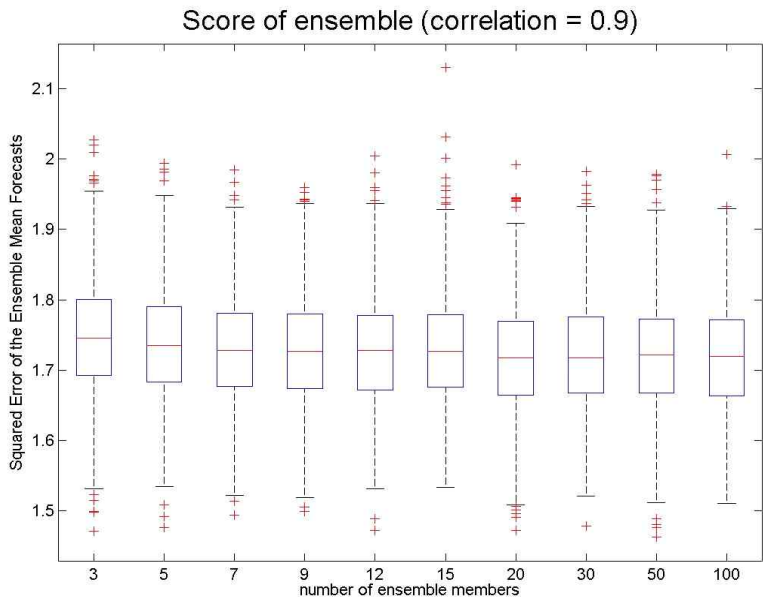


Fig. 4.34 SE of the ensemble mean forecasts (ρ_e =0.9, ρ_a = 0.1)

② $\rho_a = 0.3$

그림 4.35부터 4.41은 ρ_a가 0.3인 경우의 SE 결과를 boxplot으로 나타낸 것이다. 이는 ρ_a가 0.1인 경우에 비하여 조금 더 정확한 예측을 하는 시나리오를 가지는 앙상블 예측을 수행하는 상황이다. 따라서 ρ_a가 0.1인 경우보다 모든 경우에서 SE가 작게 나타난 것을 확인할 수 있다.

일정한 시나리오 간 상관성을 가지는 경우, 시나리오 수가 증가할수록 오차가 작아지는 것을 볼 수 있다. 그리고 시나리오 간 상관성이 점차 커질수록 오차가 줄어드는 정도가 감소하고 있다. 또한 전체적인 오차의 크기가 증가하고 있음을 확인할 수 있다. ρ_a가 0.1인 경우와 마찬가지로 0.3인 경우도 비교적 부정확한 예측을 수행하는 시나리오를 가지는 앙상블이라고 볼 수 있다. 이와 같은 경우 시나리오 간 상관성의 증가가 정보의 손실을 야기하여 오차를 증가시키는 것으로 볼 수 있다.

ρ_a가 0.3인 경우에서도 시나리오 간 상관성이 커지면서 box의 크기가 커지는 것을 확인할 수 있다. 시나리오의 상관성이 오차를 증가시킬 뿐만 아니라 오차의 분산까지 증가시키는 것을 확인할 수 있다. 또한 ρ_a가 0.1인 경우와 마찬가지로 상관성의 증가에 따라서 시나리오 수 증가의 오차 감소 효과가 작아진다. 하지만 ρ_a가 0.1인 경우에 비하여 더 작은 ρ_e에서 시나리오 수 증가가 오차 감소에 영향을 미치지 못하는 결과가 나타난다(그림 4.40).

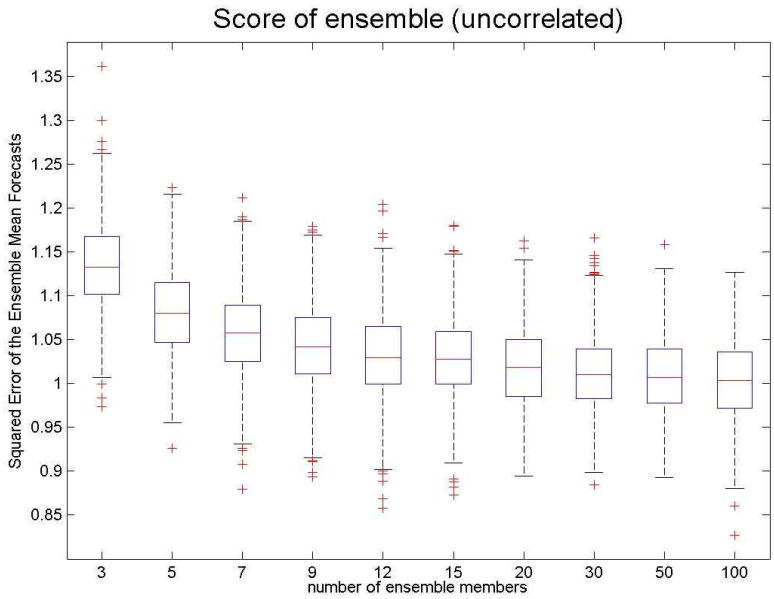


Fig. 4.35 SE of the ensemble mean forecasts (ρ_e =0, ρ_a = 0.3)

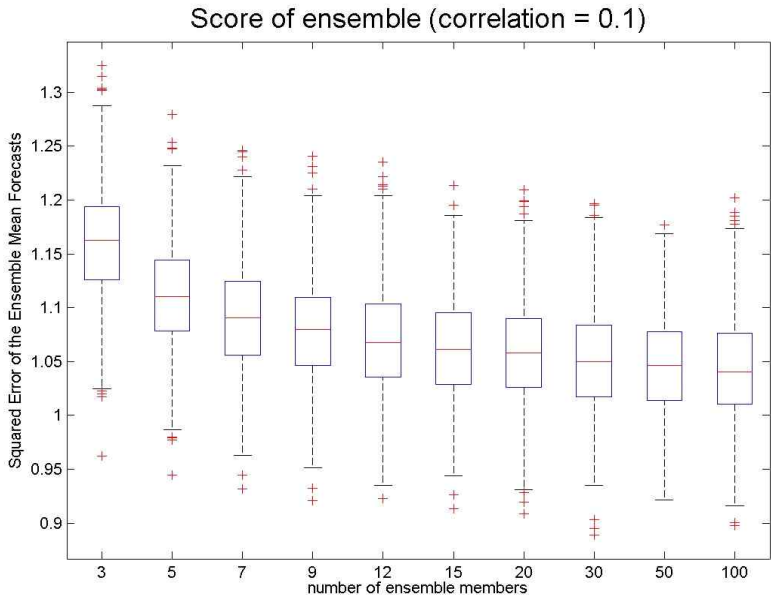


Fig. 4.37 SE of the ensemble mean forecasts (ρ_e =0.1, ρ_a = 0.3)

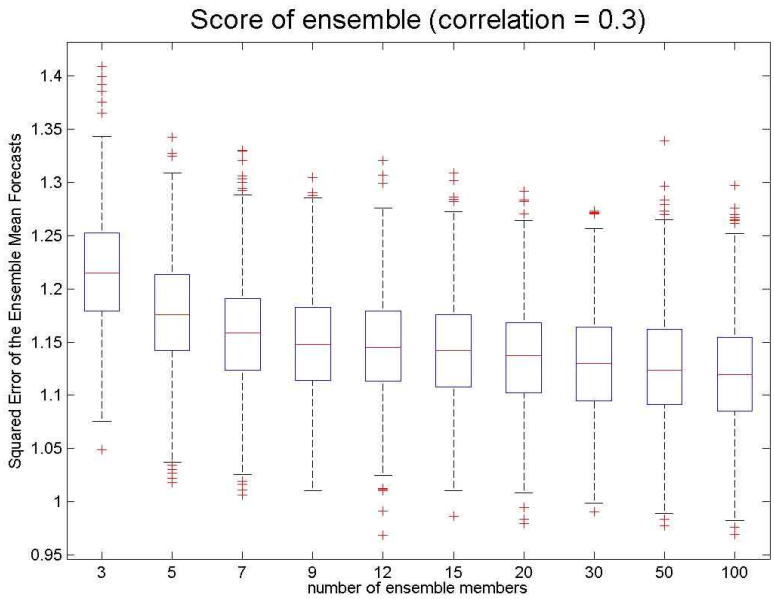


Fig. 4.38 SE of the ensemble mean forecasts (ρ_e =0.3, ρ_a = 0.3)

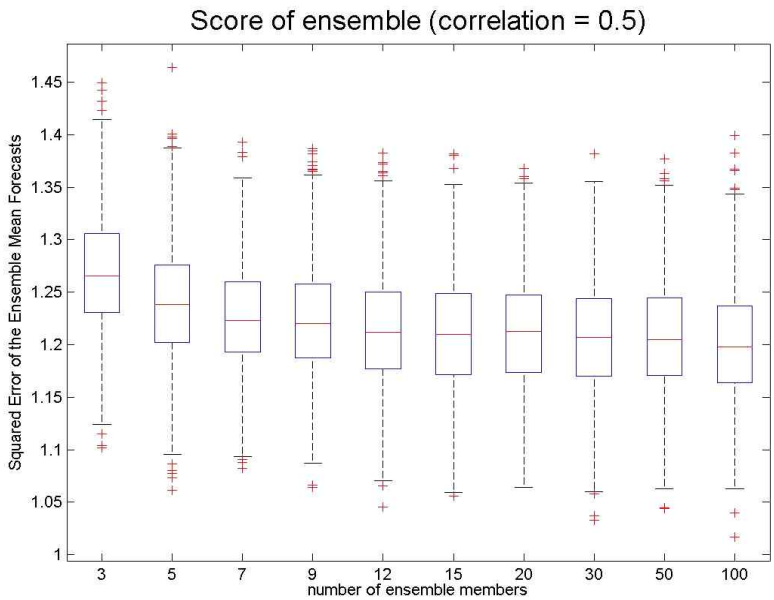


Fig. 4.39 SE of the ensemble mean forecasts (ρ_e =0.5, ρ_a = 0.3)

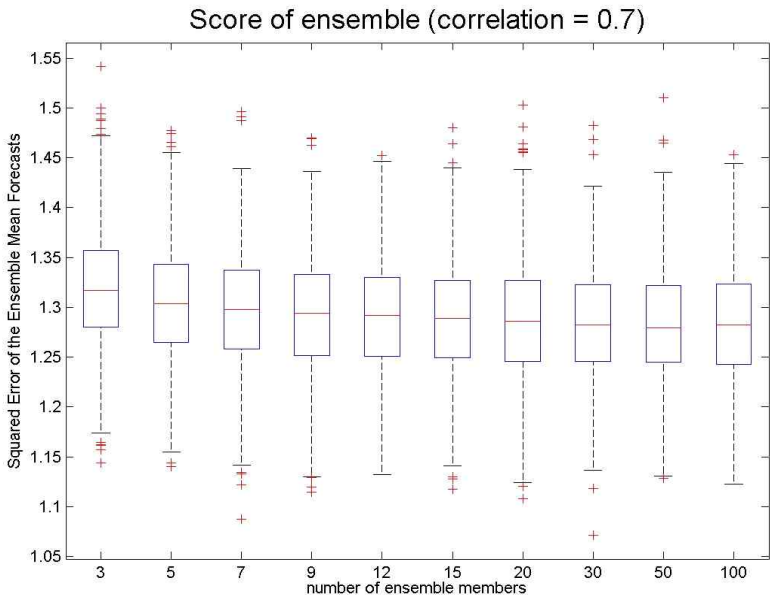


Fig. 4.40 SE of the ensemble mean forecasts (ρ_e =0.7, ρ_a = 0.3)

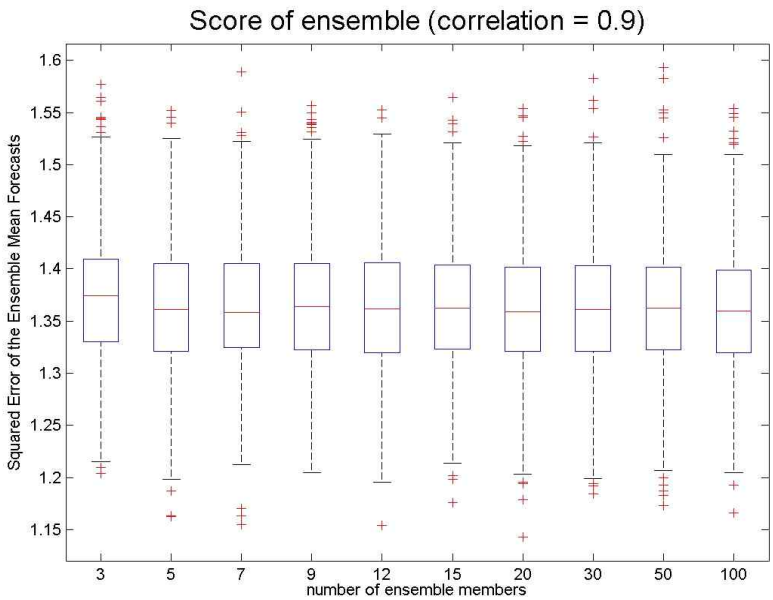


Fig. 4.41 SE of the ensemble mean forecasts (ρ_e =0.9, ρ_a = 0.3)

③ $\rho_a = 0.5$

그림 4.42부터 그림 4.47은 ρ_a가 0.5인 경우의 SE를 boxplot으로 나타낸 결과이다. 이 경우는 ρ_a가 0.1와 0.3인 경우에 비하여 더 높은 정확도를 가지는 예측을 하는 시나리오가 앙상블에 존재하는 경우를 모의한 것이다. 앞서 나타난 결과와 다른 양상을 보여준다.

앙상블 시나리오 간 상관성(ρ_e)이 0에서 0.9까지 변하더라도 같은 결과가 나타난다. 상관성 변화에 따라 오차의 크기와 분산이 변하지 않는다. 시나리오 간 상관성이 앙상블 시스템에 어떤 영향도 끼치지 못하고 있는 모습을 보이고 있다. 또한 x축 시나리오 수의 변화에 따라 SE의 증가 혹은 감소의 경향이 나타나지 않는다.

모든 경우에서 SE의 중앙값이 거의 1인 것을 확인할 수 있다. 이는 4.3.1 (1) 전체 결과 분석에서 나타났던 SE의 boxplot 결과와 비슷한 결과라는 것을 알 수 있다. 전체 결과는 SE의 분산이 앙상블 시나리오 간 상관성과 시나리오 수에 따라 변하였다. 하지만 ρ_a가 0.5인 경우에는 SE의 중앙값이 1로 나타난 결과는 같으나 분산이 항상 일정하다는 차이를 보인다.

ρ_a가 0.5라는 것이 큰 의미를 가짐으로써 시나리오 수와 시나리오 간 상관성이 SE에 영향을 미치지 못하는 것이라고 보기 어렵다. ρ_a가 0.1에서 0.5까지 증가하면서 전체적으로 오차가 줄어들며, 시나리오 수 증가의 효과가 감소하여 그래프의 box간의 기울기가 점차 완만해지는 것을 확인할 수 있다. 시나리오 수 증가의 효과와 ρ_a의 효과가 서로 상쇄된 것으로 판단된다.

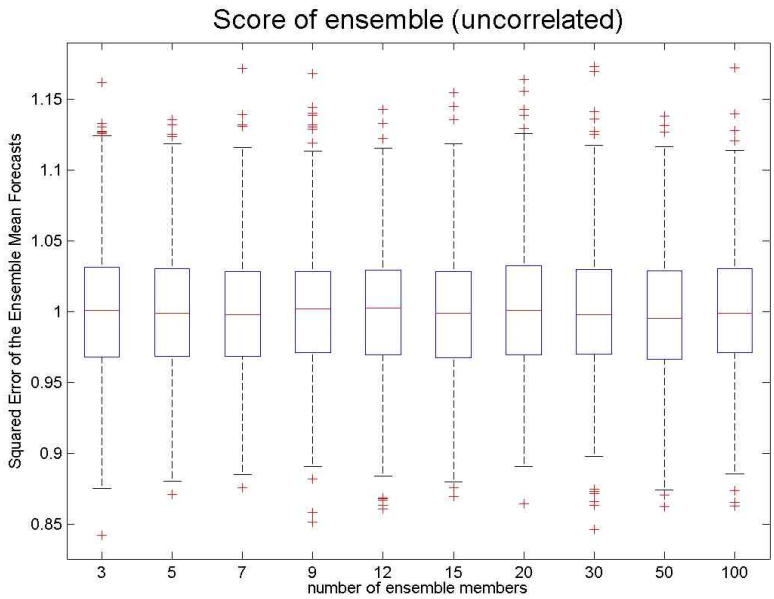


Fig. 4.42 SE of the ensemble mean forecasts (ρ_e =0, ρ_a = 0.5)

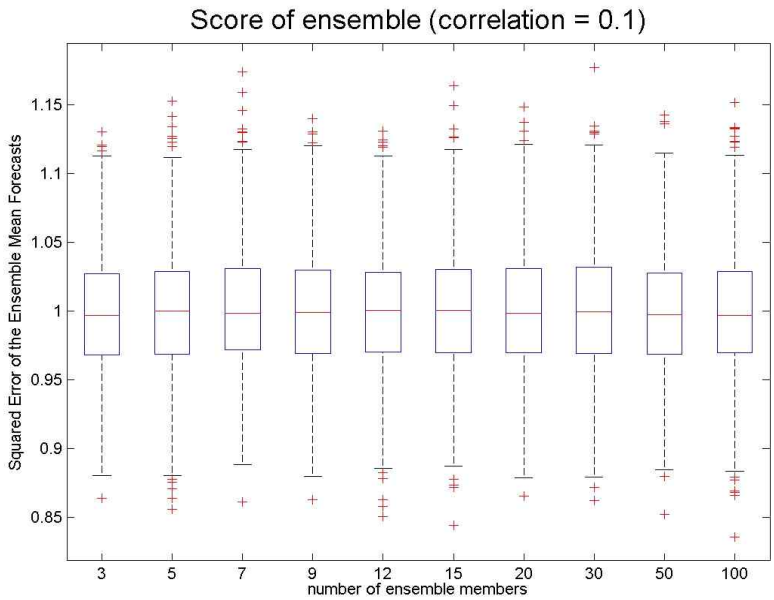


Fig. 4.43 SE of the ensemble mean forecasts (ρ_e =0.1, ρ_a = 0.5)

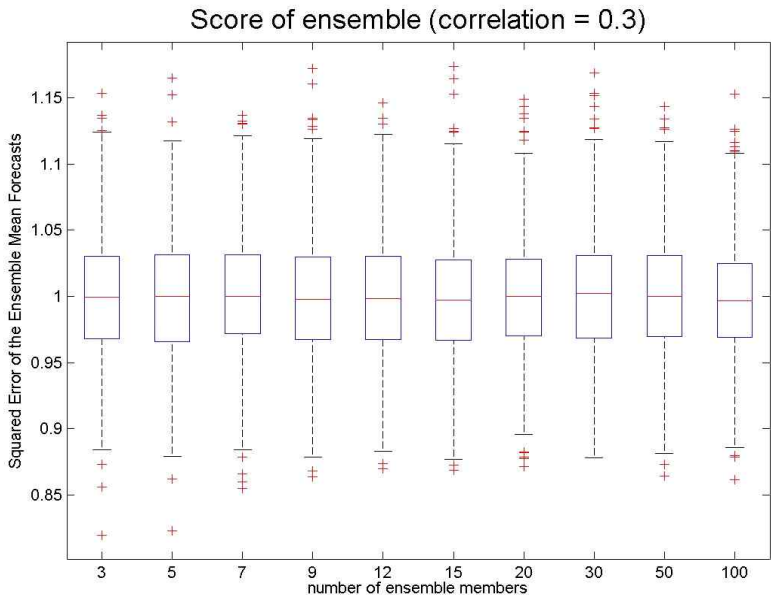


Fig. 4.44 SE of the ensemble mean forecasts (ρ_e =0.3, ρ_a = 0.5)

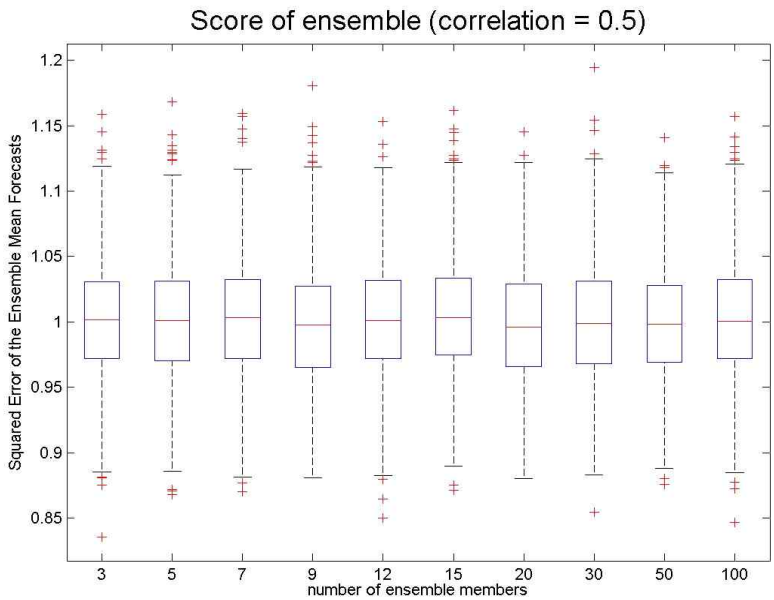


Fig. 4.45 SE of the ensemble mean forecasts (ρ_e =0.5, ρ_a = 0.5)

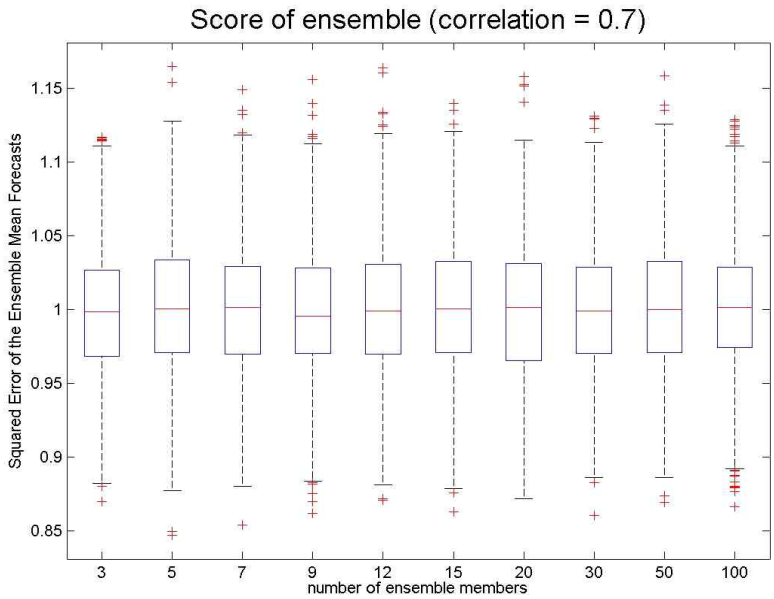


Fig. 4.46 SE of the ensemble mean forecasts (ρ_e =0.7, ρ_a = 0.5)

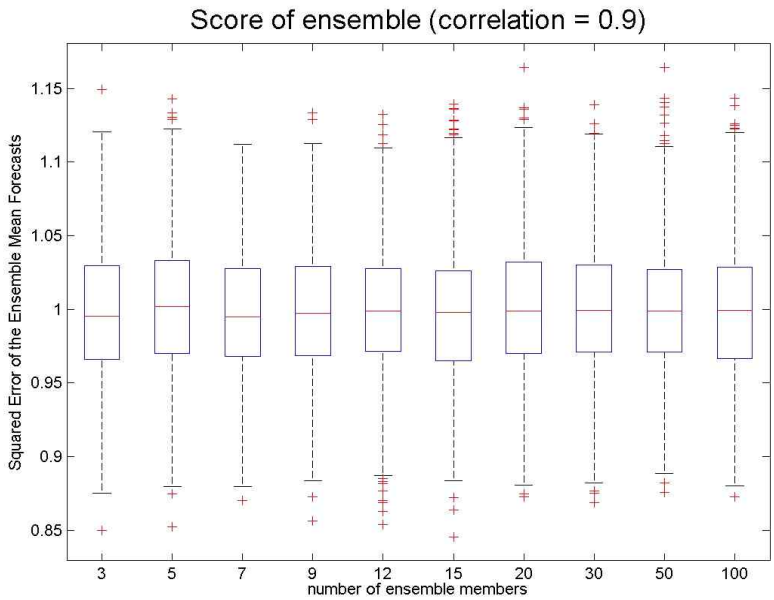


Fig. 4.47 SE of the ensemble mean forecasts (ρ_e =0.9, ρ_a = 0.5)

④ $\rho_a = 0.7$

그림 4.49부터 그림 4.54는 ρ_a가 0.7일 때의 SE 결과를 boxplot으로 도시한 것이다. 이 경우는 비교적 높은 정확도를 가지는 예측하는 시나리오가 앙상블 내에 반드시 하나가 존재하고 있다. 앞선 결과들과 비교하였을 때 모든 ρ_e의 경우에서 낮은 오차를 보이고 있으며 항상 1보다 낮은 SE 결과를 보여준다.

앙상블 시나리오 간 상관성(ρ_e)이 커질수록 SE가 작아지는 결과가 나타난다. 이전 ρ_a가 0.1, 0.3 일때는 시나리오 간 상관성 증가로 인하여 예측 시스템이 가지는 정보의 감소를 야기하고, 오차가 증가하였다. 하지만 정확도가 높은 시나리오가 있는 경우 시나리오 간 상관성(ρ_e)이 커질수록 이 시나리오와 비슷한 수준의 예측을 하는 시나리오들이 증가하게 된다.

시나리오 수가 앙상블 예측 시스템에 미치는 양상도 ρ_a가 작을 때와 비교하여 달라지는 것을 확인할 수 있다. 시나리오 수 증가가 오히려 평균 앙상블 예측의 정확도를 낮게 만드는 결과를 야기한다. 이 경향성은 시나리오 간 상관성(ρ_e)이 다른 값을 가져도 변하지 않는다. 높은 정확도를 가지는 시나리오($\boldsymbol{y_1}$)가 있는 앙상블인 경우, 다른 시나리오 수가 적을수록 정확도 높은 시나리오의 영향이 커지게 되어 평균 앙상블의 오차가 작아지게 된다.

시나리오 수가 증가할수록 오차가 커지는 정도가 ρ_e가 증가할수록 작아지는 것을 확인할 수 있다. ρ_e가 증가할수록 정확도 높은 시나리오($\boldsymbol{y_1}$)와 비슷한 퍼포먼스를 보이는 시나리오가 증가하게 된다. 따라서 정확도 높은 시나리오의 효과를 덜 감소시켜 이와 같은 결과가 나타난다.

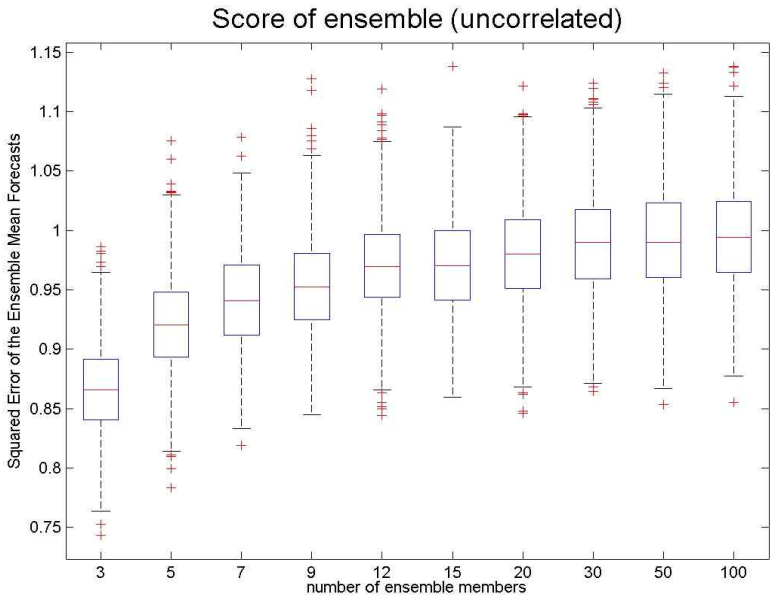


Fig. 4.48 SE of the ensemble mean forecasts (ρ_e =0, ρ_a = 0.7)

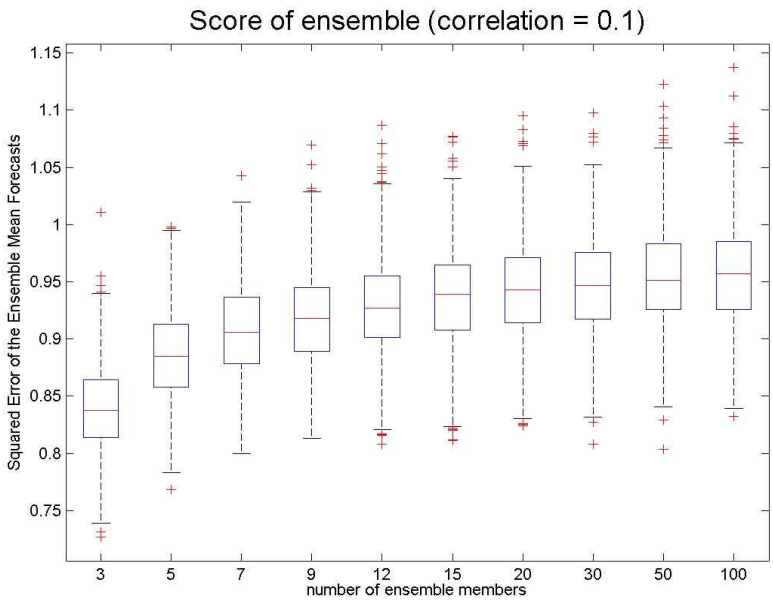


Fig. 4.49 SE of the ensemble mean forecasts (ρ_e =0.1, ρ_a = 0.7)

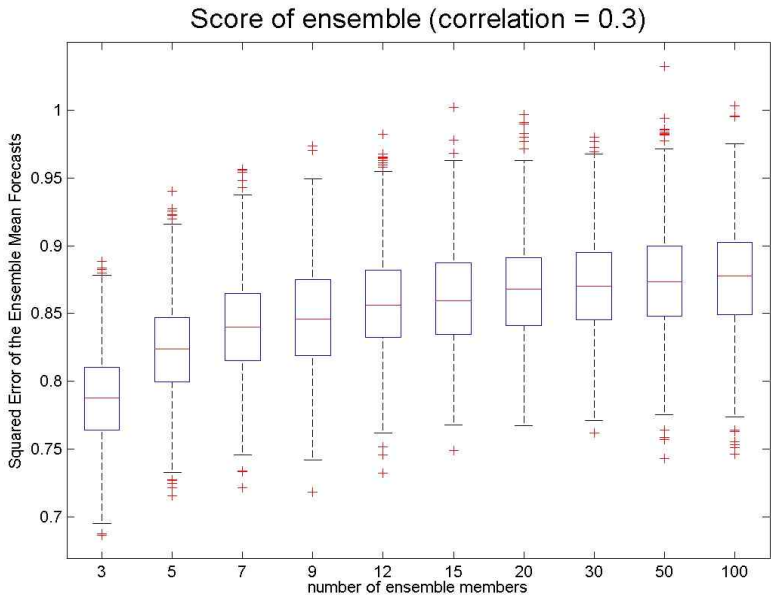


Fig. 4.50 SE of the ensemble mean forecasts (ρ_e =0.3, ρ_a = 0.7)

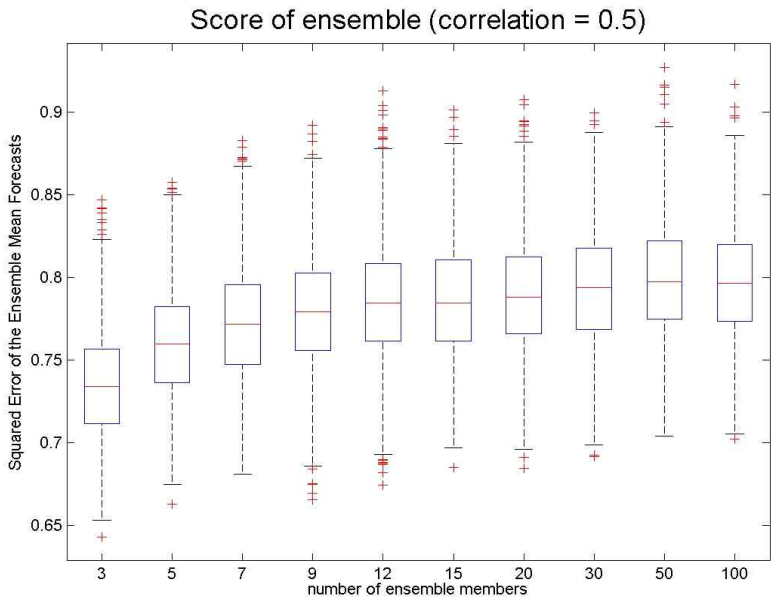


Fig. 4.51 SE of the ensemble mean forecasts (ρ_e =0.5, ρ_a = 0.7)

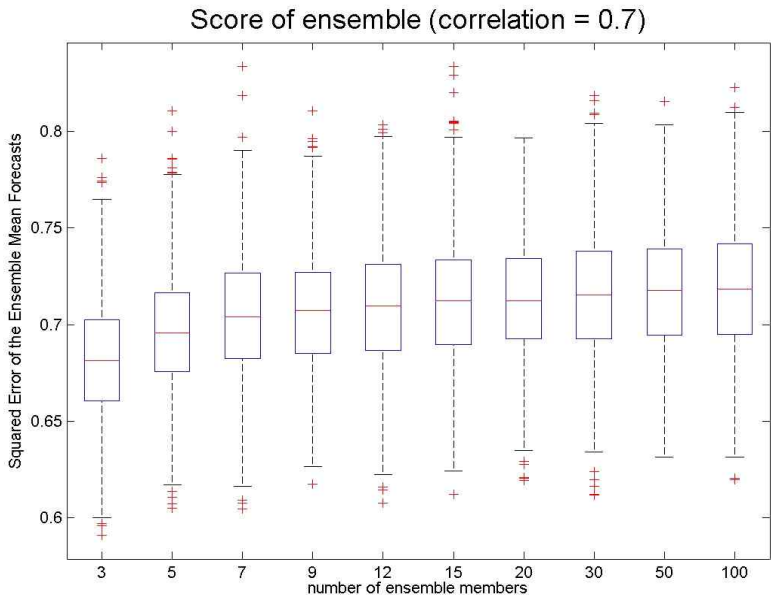


Fig. 4.52 SE of the ensemble mean forecasts (ρ_e =0.7, ρ_a = 0.7)

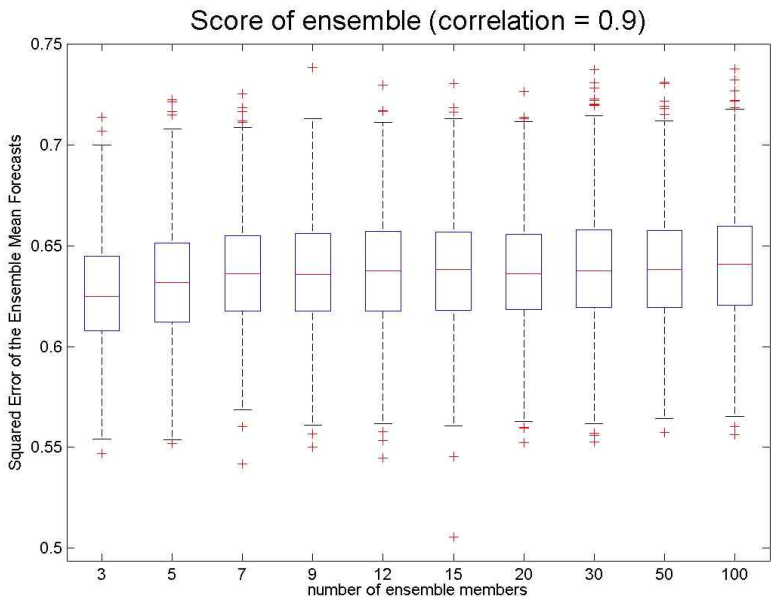


Fig. 4.53 SE of the ensemble mean forecasts (ρ_e =0.9, ρ_a = 0.7)

⑤ $\rho_a = 0.9$

그림 4.54부터 그림 4.59는 ρ_a가 0.9일 때 결과를 ρ_e별로 boxplot으로 나타낸 것이다. 앞의 모든 경우 중 가장 정확도 높은 시나리오를 가지는 앙상블 예측 시스템이다. 그러므로 모든 앙상블 시나리오 간 상관성(ρ_e)의 경우에서 SE가 전체적으로 가장 낮게 나타난다. 정확한 예측을 수행할 수 있는 앙상블에 시나리오 간 상관성이 있을 때의 평균 앙상블 예측의 오차를 살펴볼 수 있다.

이 경우도 ρ_a가 0.7인 경우와 마찬가지로 ρ_e가 증가할수록 SE가 작아지는 결과가 나타난다. ρ_e가 증가함에 따라 앙상블 내의 시나리오들이 전반적으로 정확도가 올라가서 생기는 결과라고 할 수 있다. 시나리오 간 상관성이 증가하여 잃은 정보의 손해보다 정확한 예측으로 얻는 오차의 감소가 더 큰 경우라고 할 수 있다.

ρ_a가 0.7일 때와 마찬가지로 이 경우에도 앙상블 시나리오 수가 증가할수록 오차가 커지는 것을 확인할 수 있다. $\boldsymbol{y_1}$의 정확도가 올라감으로써 다른 시나리오 수의 증가가 예측의 정확도를 낮추는 결과를 보인다. 그리고 시나리오 수 대비 정확도 감소의 효과가 ρ_a가 0.7인 경우와 비교해서 더 강하게 나타나고 있다. 이와 같이 아주 정확한 예측을 수행할 수 있는 경우에는 앙상블 시나리오 간 상관성 증가가 반드시 좋은 효과를 야기하는 것이 아니라는 것을 의미한다.

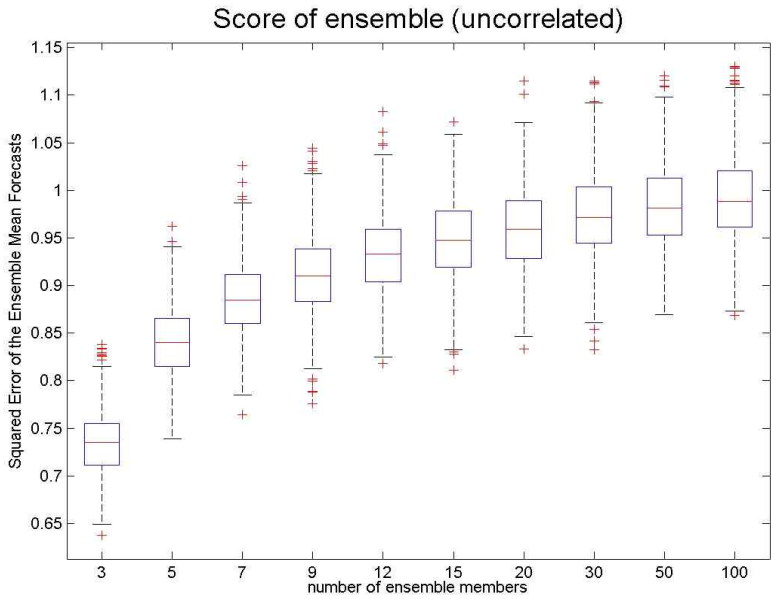


Fig. 4.54 SE of the ensemble mean forecasts (ρ_e =0, ρ_a = 0.9)

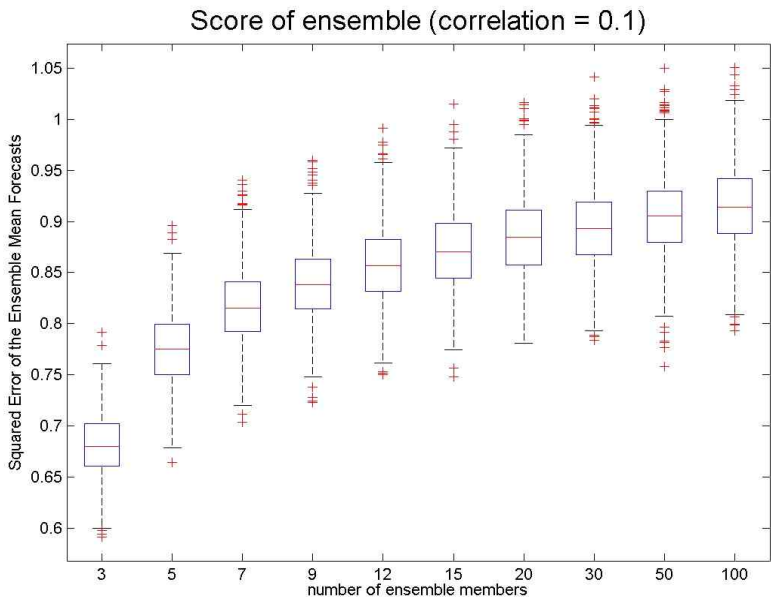


Fig. 4.55 SE of the ensemble mean forecasts (ρ_e =0.1, ρ_a = 0.9)

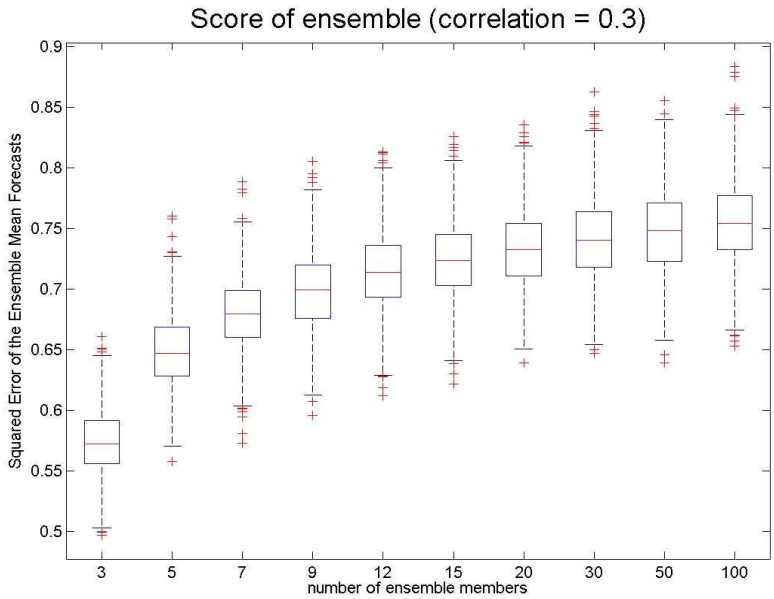


Fig. 4.56 SE of the ensemble mean forecasts (ρ_e =0.3, ρ_a = 0.9)

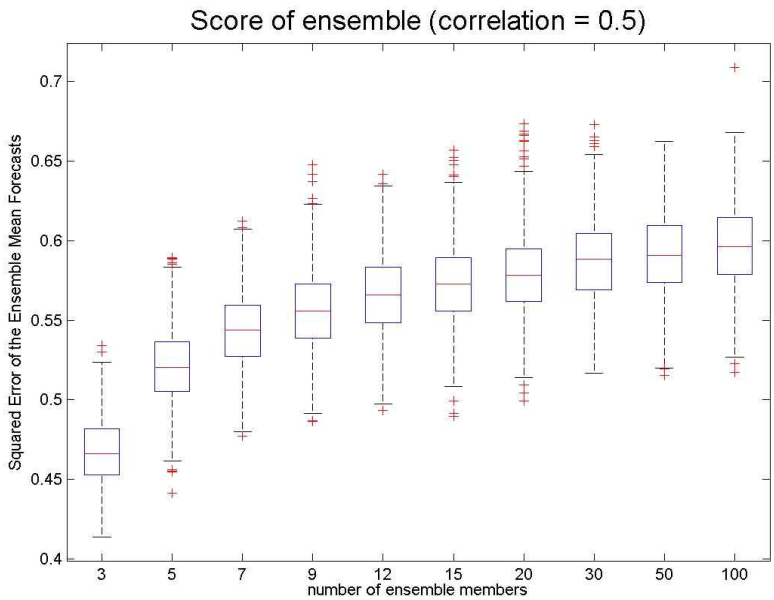


Fig. 4.57 SE of the ensemble mean forecasts (ρ_e =0.5, ρ_a = 0.9)

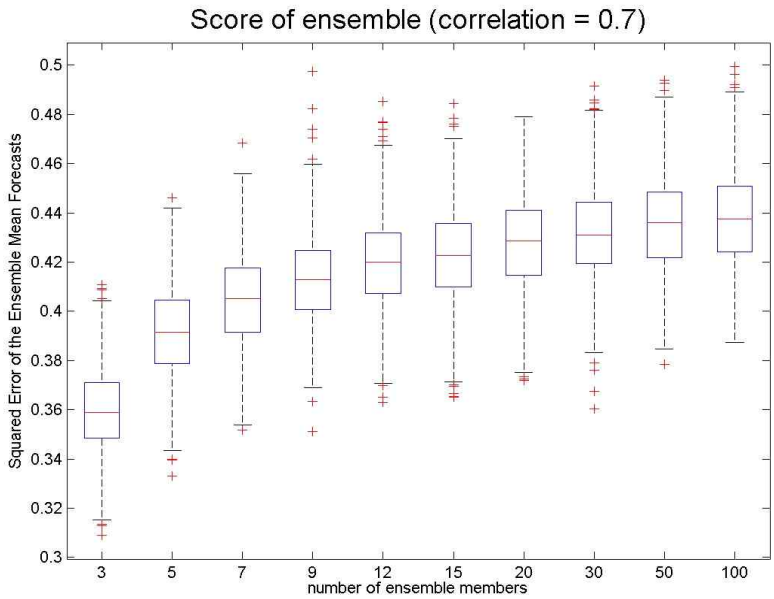


Fig. 4.58 SE of the ensemble mean forecasts (ρ_e =0.7, ρ_a = 0.9)

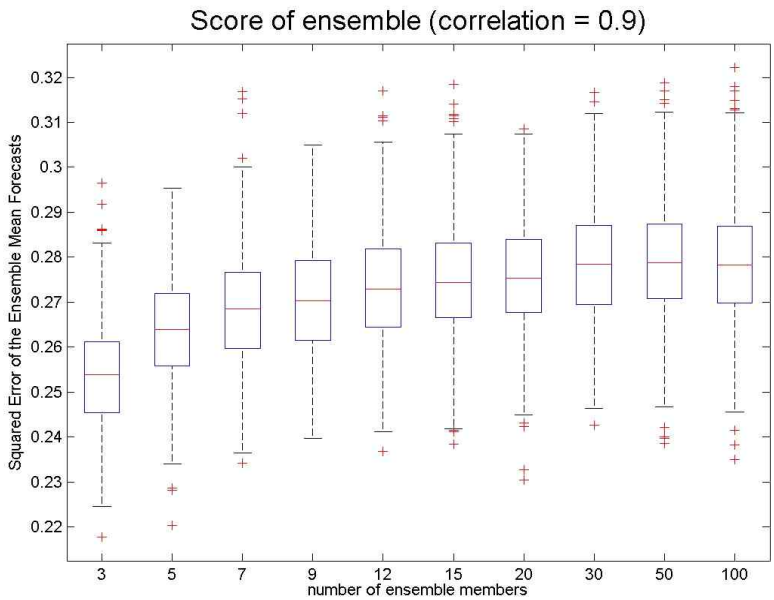


Fig. 4.59 SE of the ensemble mean forecasts (ρ_e =0.9, ρ_a = 0.9)

4.3.2 다중범주 확률예측 평가

(1) 전체 결과 분석

앙상블 시나리오 간 상관성(ρ_e)의 효과를 분석하기 위하여 다중범주 확률예측 평가지표로서 BS를 사용하였다. 앙상블 예측은 한 시점에 대하여 여러 가지 가능한 예측값을 제공하는 확률예측방법이다. 따라서 앙상블 예측을 평가하는데 있어서 다중범주 확률예측 평가기법이 확정적 예측 기법보다 앙상블 전체의 예측 능력을 평가하는데 더 적합하다고 볼 수 있다. 앞에서 사용한 SE는 평균앙상블 시나리오의 거동을 보여준 지표이며, 이 절에서는 Brier Score를 이용하여 전체 앙상블의 정확도를 평가하였다.

그림 4.60부터 그림 4.65은 시나리오 간 상관성이 앙상블 예측 시스템에 미치는 영향을 BS를 사용하여 분석한 결과를 나타낸 것이다. 시나리오 간 상관성의 영향만을 확인하기 위하여 ρ_a별 결과 중 같은 시나리오 간 상관계수를 가지는 결과를 이용하여 각각의 그래프를 그렸다. x축은 앙상블 시나리오 수를 나타내며 y축은 BS를 나타낸다.

그림 4.60은 앙상블 시나리오 간 상관성(ρ_e)이 0일 경우의 BS결과를 보여준다. 시나리오 수가 증가할수록 BS의 분산이 점차 작아지고 있으며 평균이 감소하고 있다. BS는 0부터 2사이의 값을 가지며 0에 가까울수록 좋은 평가를 받는 점수이다. 따라서 시나리오 간 상관성이 없을 때 시나리오 수가 증가할수록 앙상블 시스템이 더 좋은 예측을 수행할 가능성이 높아진다는 것을 알 수 있다.

Boxplot을 살펴보면 시나리오 수가 작은 경우의 최소값이 시나리오 수가 많은 경우보다 작은 것을 볼 수 있다. 이는 시나리오 수가 작을 때 더 좋은 예측을 수행할 수도 있다는걸 의미한다. 하지만 확률적으로 접근을 하면 시나리오 수가 50개 이상에서는 항상 BS가

0.7 이하로 나타나지만 9개 이하부터는 50%의 확률도 되지 못한다고 해석할 수 있다.

시나리오 간 상관성(ρ_e)이 증가하면 전반적으로 BS 평균이 커지는 것을 확인할 수 있다(그림 4.61～그림 4.65). 또한 boxplot의 크기가 커져 분산이 증가한다. 시나리오 간 상관성 증가로 인하여 정보의 손실이 발생하고, 앙상블 시스템이 가능한 예측의 폭이 줄어들게 된다. 이로 인하여 앙상블이 더 극단적인 예측을 수행하게 됨을 알 수 있다. 많은 수의 시나리오를 사용함으로써 이러한 극단적인 예측을 방지하게 된다. 하지만 ρ_e가 증가함으로써 시나리오 수 증가로 인한 효과가 감소하여 boxplot 간 크기의 차이가 거의 나타나지 않으며(그림 4.65), 시나리오 수 대비 boxplot의 평균의 차이가 점차 감소한다.

모든 ρ_e에 관하여 결과를 정리하면 그림 4.66과 같다. 한 그래프에 표현하기 위하여 각각의 boxplot을 평균하여 나타내었다. 모든 ρ_e의 경우에서 시나리오 수가 증가하면서 BS가 작아지는 것을 확인할 수 있다. ρ_a별로 분해하지 않은 결과에서는 ρ_e가 작을수록 더 낮은 SE를 평균적으로 가지게 된다. 즉 앙상블 예측 시스템 내에 시나리오가 어느 정도의 정확도를 가지는지 알 수 없을 경우, 시나리오간 상관성이 작을수록 더 좋은 예측을 할 확률이 증가하게 된다.

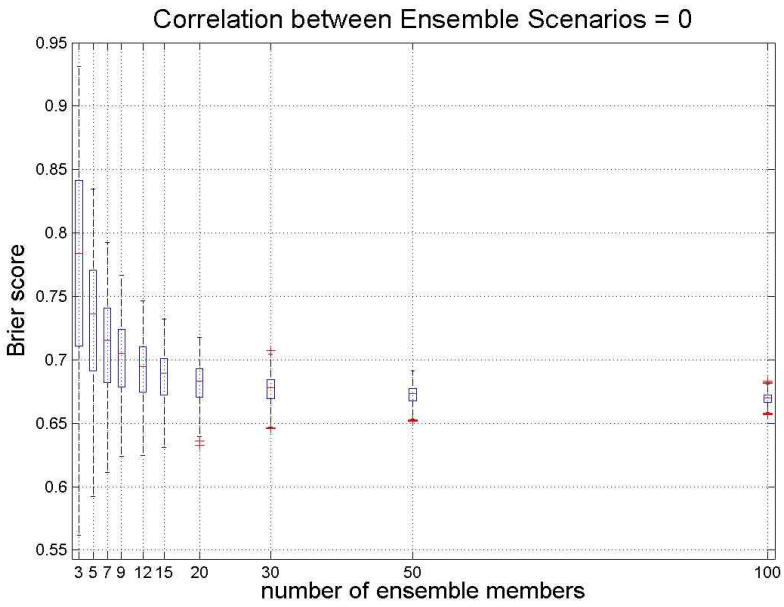


Fig. 4.60 Brier score when $\rho_e = 0$

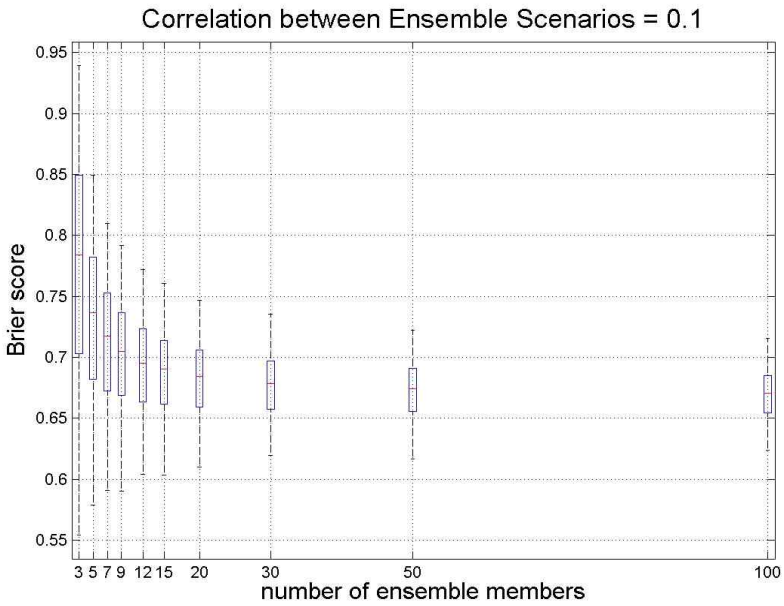


Fig. 4.61 Brier score when $\rho_e = 0.1$

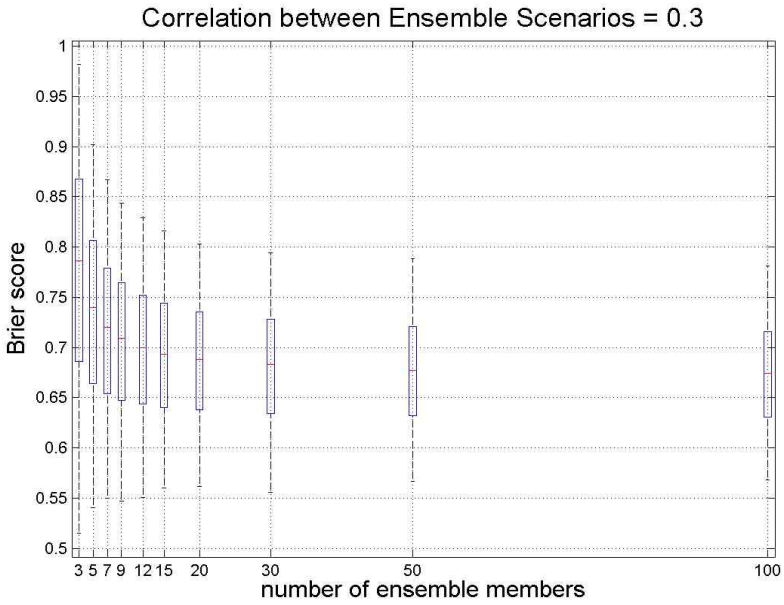


Fig. 4.62 Brier score when $\rho_e = 0.3$

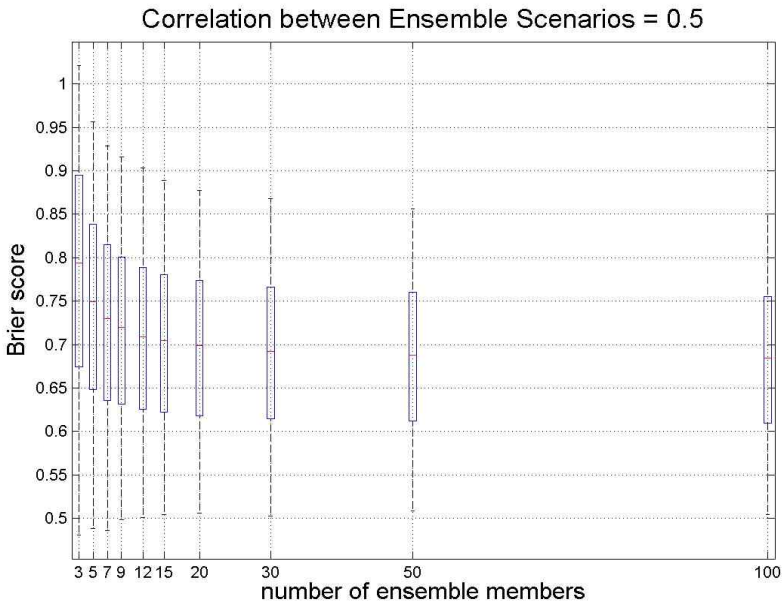


Fig. 4.63 Brier score when $\rho_e = 0.5$

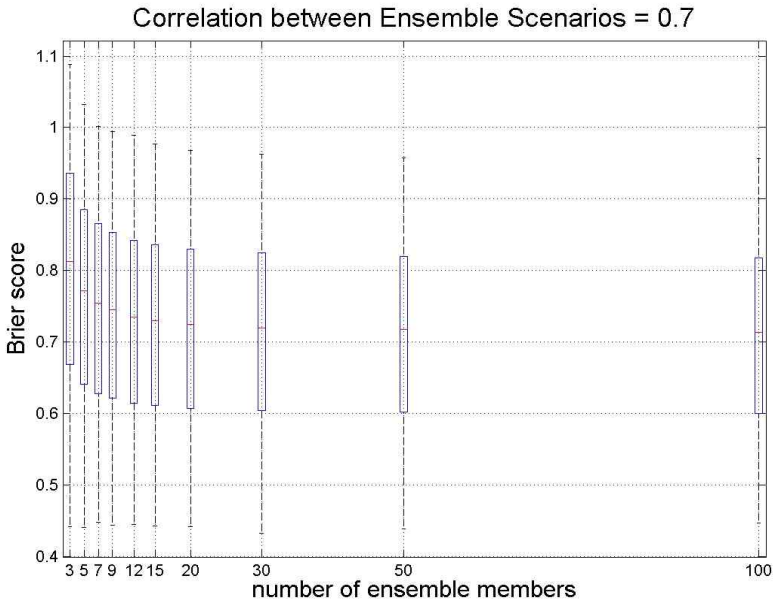


Fig. 4.64 Brier score when $\rho_e = 0.7$

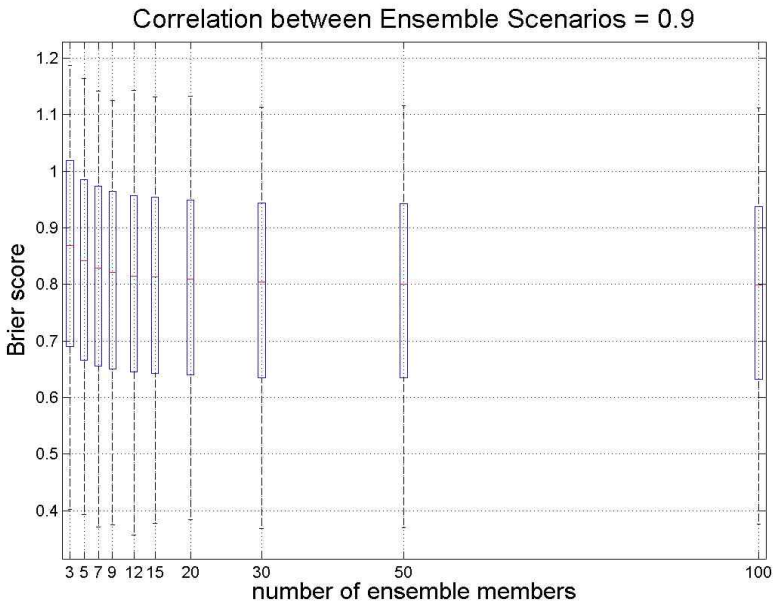


Fig. 4.65 Brier score when $\rho_e = 0.9$

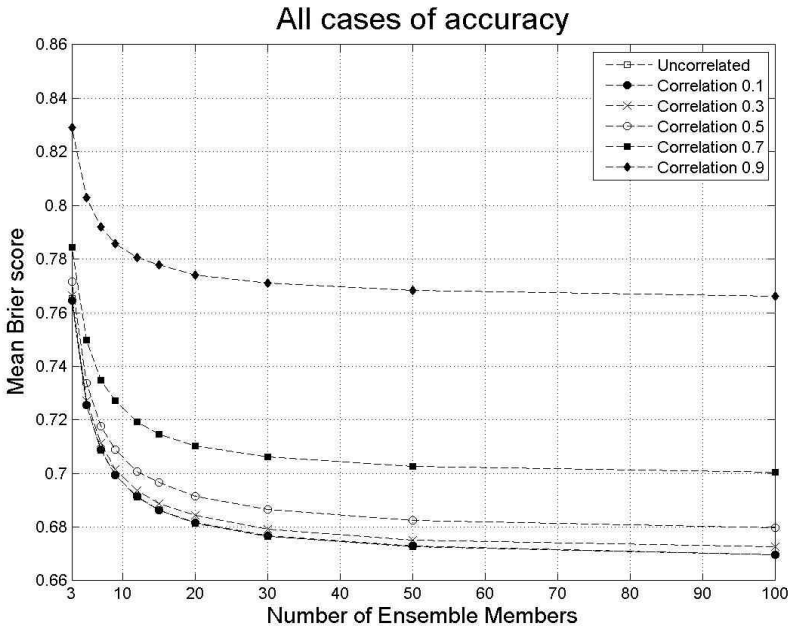


Fig. 4.66 Summary of Brier score results

(2) ρ_a별 결과 분석

본 절에서는 관측시나리오 생성 시에 사용했던 시나리오 y_1과 관측시나리오 간의 상관계수인 ρ_a별로 결과를 나타내었다. 앞 절에서 사용한 결과는 ρ_a별 결과를 앙상블 시나리오 간 상관계수인 ρ_e별로 분류하여 나타낸 것이다. 따라서 전체 분석 결과는 앙상블 예측 시스템이 얼마나 정확한 예측을 수행할 수 있는지 알 수 없는 경우이며, ρ_a별 분석 결과는 앙상블 내에 시나리오가 어느 정도 정확도를 가지는지 알 수 있는 경우를 나타낸다.

그림 4.67은 그림 4.66을 ρ_a로 분해한 결과이다. 그림 4.66에서 각각의 ρ_e별 그래프를 앙상블 내의 시나리오 정확도를 간접적으로 나타내는 ρ_a별로 분해하여 도시하면 그림 4.67과 같다. 여기에 나타난 바와 같이 동일한 ρ_e를 가지더라도 ρ_a에 의해서 아주 다른 거동을 보일 수 있다. 더 정확도가 높은 예측을 수행할 수 있더라도 시나리오 간 상관성에 의해서 BS 분석결과가 더 낮게 나타나는 경우도 확인할 수 있다.

본 절에서는 ρ_a별로 결과를 분석하였다. ρ_a가 0.1, 0.3, 그리고 0.5인 경우는 비교적 정확도가 낮은 예측을 수행하는 경우이며, 0.9인 경우는 높은 정확도의 예측을 수행할 수 있는 경우를 의미한다. 각 경우마다 시나리오 간 상관성과 시나리오 수가 앙상블 예측 시스템에 미치는 영향을 BS 지표를 이용하여 분석하였다. 앞서 사용한 평균 시나리오의 정확도를 평가하는데 사용한 SE 지표와는 달리 BS를 이용하여 확률 예측으로서의 앙상블을 평가하는데 목적이 있다.

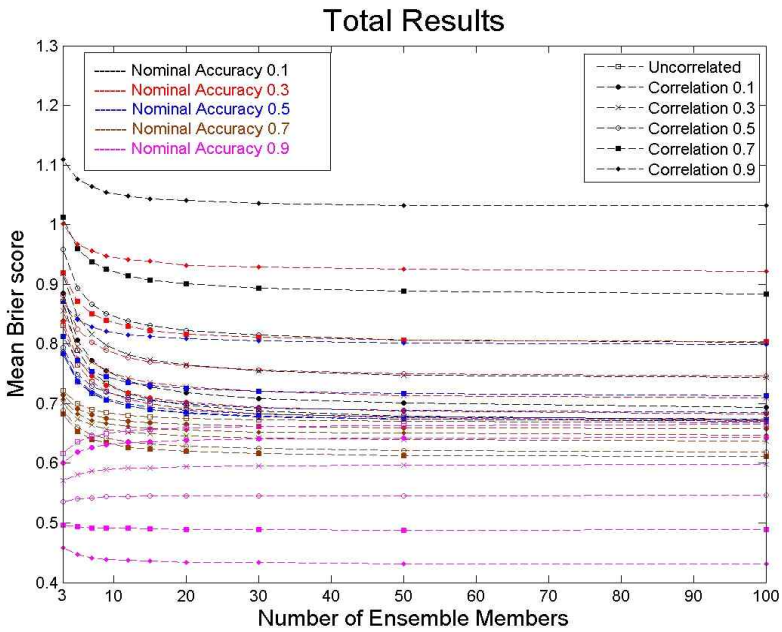


Fig. 4.67 Summary of Brier score results(by nominal accuracy, ρ_a)

① $\rho_a = 0.1$

그림 4.68부터 그림 4.73은 ρ_a가 0.1인 경우의 BS 결과를 boxplot으로 나타낸 것이다. x축은 앙상블 내 시나리오 수를 나타내며, y축은 BS를 나타낸다. 그리고 시나리오 간 상관성(ρ_e)별로 결과를 분류하였다. 그림 4.68은 시나리오 $\boldsymbol{y_1}$의 정확도 ρ_a가 0.1이며 ρ_e가 0인 경우이다. 앙상블 예측 시스템이 비교적 부정확한 예측을 수행하며, 앙상블 내 시나리오 간 상관성이 없는 경우를 나타낸다. 이 경우 시나리오 수가 증가하면 BS가 작아진다. 즉, 많은 시나리오를 사용할수록 예측이 향상되는 것을 확인할 수 있다.

시나리오 간 상관성(ρ_e)이 증가할수록 전체적인 오차가 커지는 것을 볼 수 있다. 이는 비교적 부정확한 예측을 할 수밖에 없는 경우에는 시나리오 간 상관성이 작을수록 앙상블 예측의 정확도가 올라간다는 것을 의미한다. 작은 시나리오 간 상관성이 예측의 범위를 넓혀 확률예측의 정확도가 향상된다고 볼 수 있다. 또한 앙상블 시나리오 간 상관성이 증가하면서 시나리오 수 증가에 따른 예측향상 정도가 감소하는 것을 확인할 수 있다. ρ_e가 0.9인 경우(그림 4.73)과 ρ_e가 0인 경우(그림 4.68)을 비교하였을 때 이러한 경향을 쉽게 확인할 수 있다. 상관성 증가로 인한 정보의 손실 효과가 시나리오 수 증가 효과에 비해 더 크게 작용하고 있는 것으로 볼 수 있다.

ρ_e가 증가할수록 시나리오 수에 따른 boxplot 크기의 차이가 감소하는 것을 볼 수 있다. ρ_e가 0일 때 시나리오 수가 100인 경우가 3인 경우보다 boxplot의 크기가 작은 것을 볼 수 있다. 하지만 ρ_e가 커지면서 시나리오 수 100과 시나리오 수 3인 경우의 boxplot의 크기 차이가 작아지는 것을 볼 수 있다. 시나리오 간 상관성(ρ_e)이 커짐에 따라 오차의 분산이 증가하므로 더 큰 오차가 발생할 확률이 증가한다.

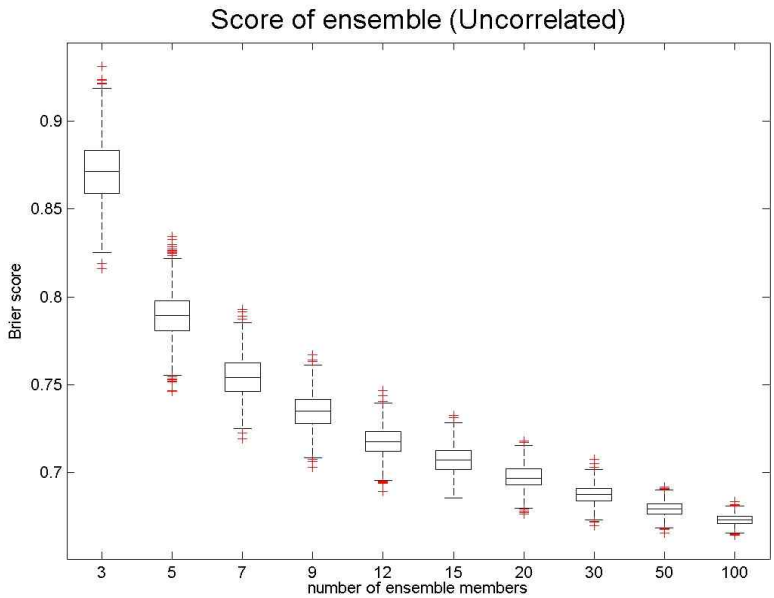


Fig. 4.68 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.1$)

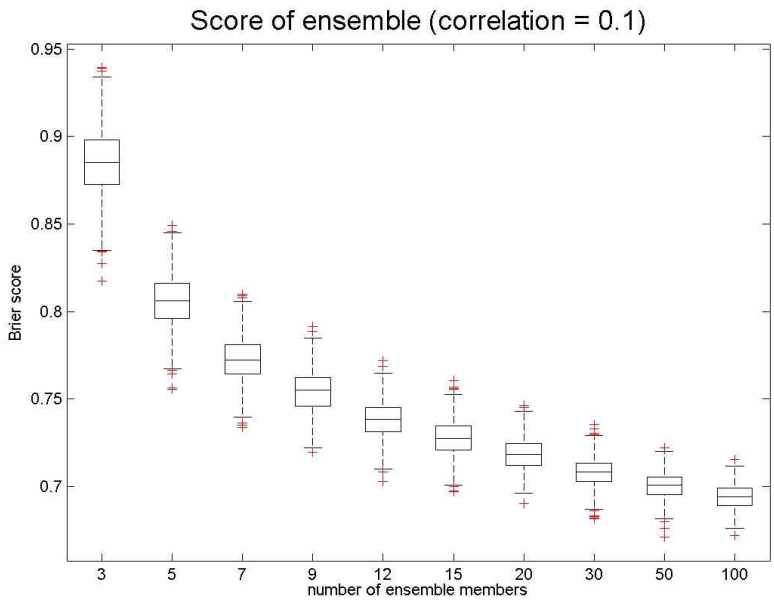


Fig. 4.69 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.1$)

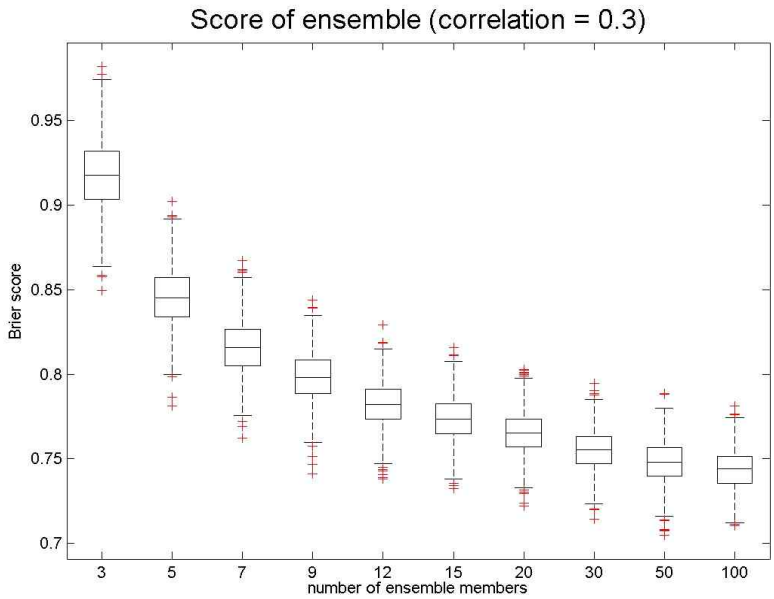


Fig. 4.70 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.1$)

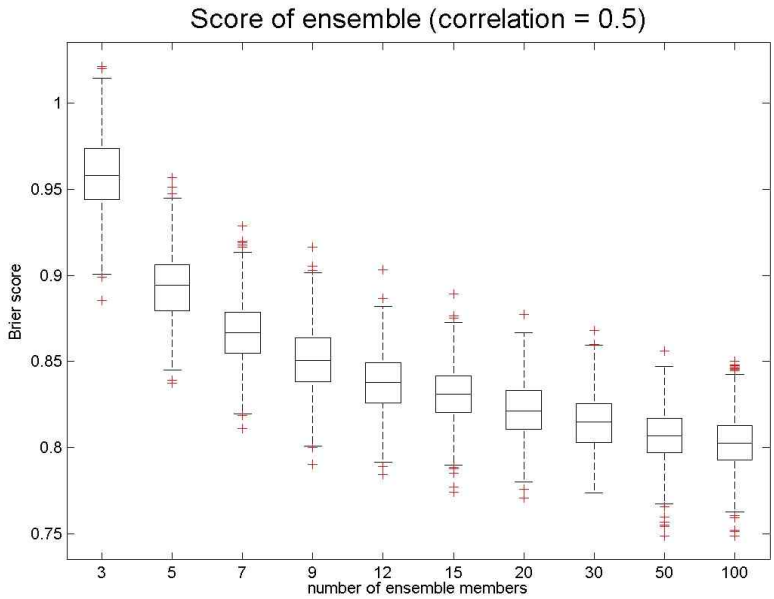


Fig. 4.71 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.1$)

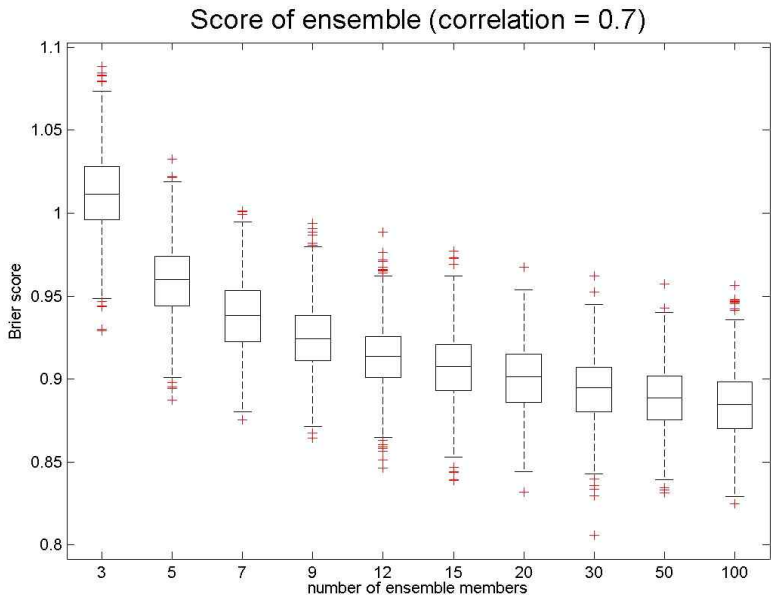


Fig. 4.72 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.1$)

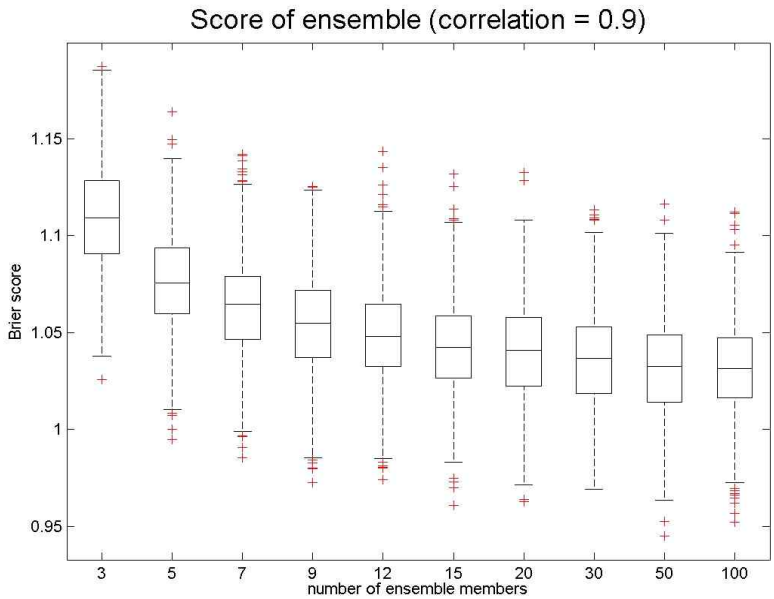


Fig. 4.73 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.1$)

② $\rho_a = 0.3$

그림 4.74부터 그림 4.79는 ρ_a가 0.3인 경우의 BS 결과를 boxplot으로 나타낸 것이다. 전반적으로 ρ_a가 0.1일 때의 결과와 비슷한 양상을 보인다. ρ_a가 0.3인 경우는 ρ_a가 0.1인 앞의 경우에 비하여 $\boldsymbol{y_1}$의 정확도가 높아 전체적으로 앙상블 예측의 정확도가 향상된 경우이다. 그러므로 ρ_a가 0.1인 경우와 비교하였을 때 모든 경우에서 BS가 작게 나타나 ρ_e와 무관하게 모든 경우에서 앙상블 예측이 향상된 것을 확인할 수 있다.

앞선 경우와 마찬가지로 시나리오 간 상관성(ρ_e)이 증가할수록 전체적인 오차가 증가하는 것을 확인할 수 있다. 이는 ρ_a가 0.3인 경우도 $\boldsymbol{y_1}$의 예측 정확도가 비교적 부정확하기 때문에 시나리오 간 상관성이 작을수록 가능한 예측의 범위가 증가하여 예측에 유리하게 작용했다고 볼 수 있다. 또한 ρ_a=0.1과 마찬가지로 앙상블 시나리오 수가 증가하면 BS가 작아지는 것을 확인할 수 있다. 앙상블 시나리오 간 상관성이 증가하면서 시나리오 수 증가에 따른 BS 향상 효과가 감소하는 것으로 보인다. 시나리오 간 상관성 증가로 인한 정보 손실이 시나리오 수 증가로 인한 예측결과 향상 효과를 감소시킨 것이라 할 수 있다.

ρ_a가 0.3인 경우에서도 시나리오 간 상관성이 커지면서 전체적인 boxplot의 크기가 커지는 것을 확인할 수 있다. 시나리오의 상관성이 오차를 증가시킬 뿐만 아니라 오차의 분산까지 증가시키는 것을 확인할 수 있다. 또한 시나리오 수가 작은 앙상블 예측의 boxplot과 시나리오 수가 많은 예측의 boxplot의 차이가 ρ_e가 커짐에 따라 감소한다. 이 경우도 시나리오 간 상관성이 작을수록 예측에 유리하게 작용한다는 것을 확인할 수 있다.

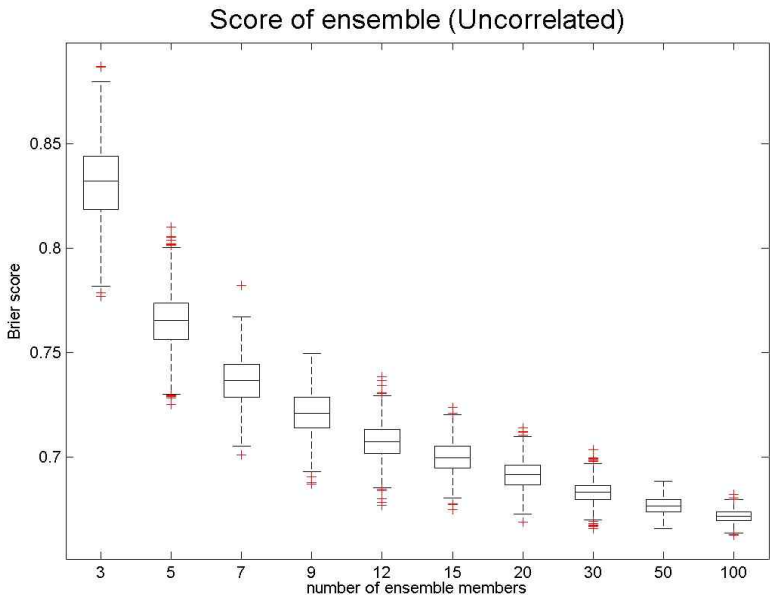


Fig. 4.74 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.3$)

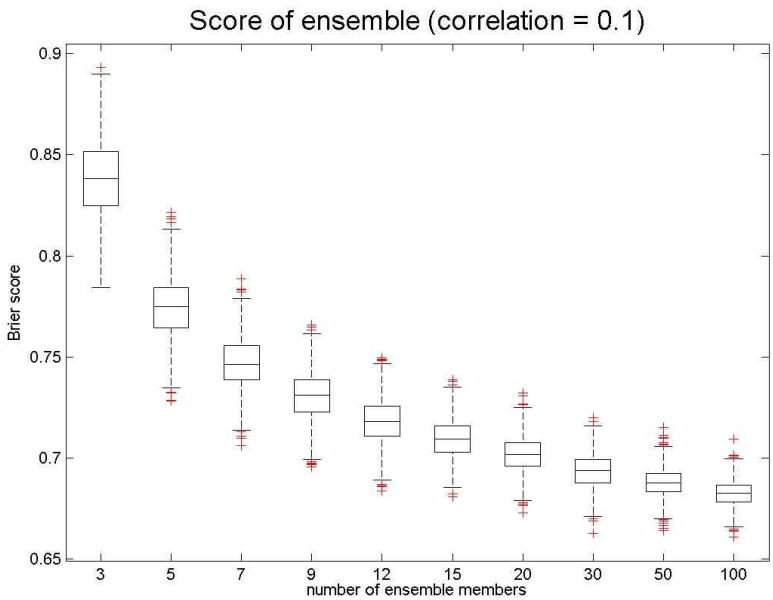


Fig. 4.75 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.3$)

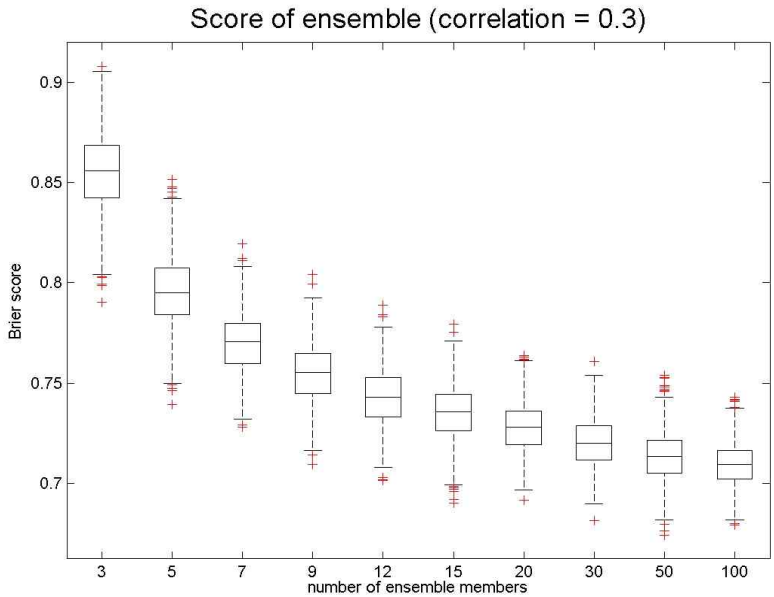


Fig. 4.76 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.3$)

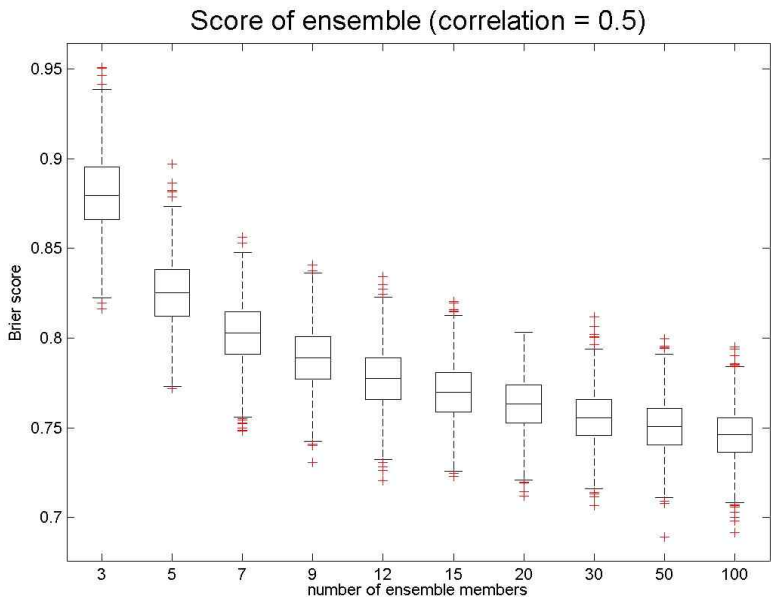


Fig. 4.77 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.3$)

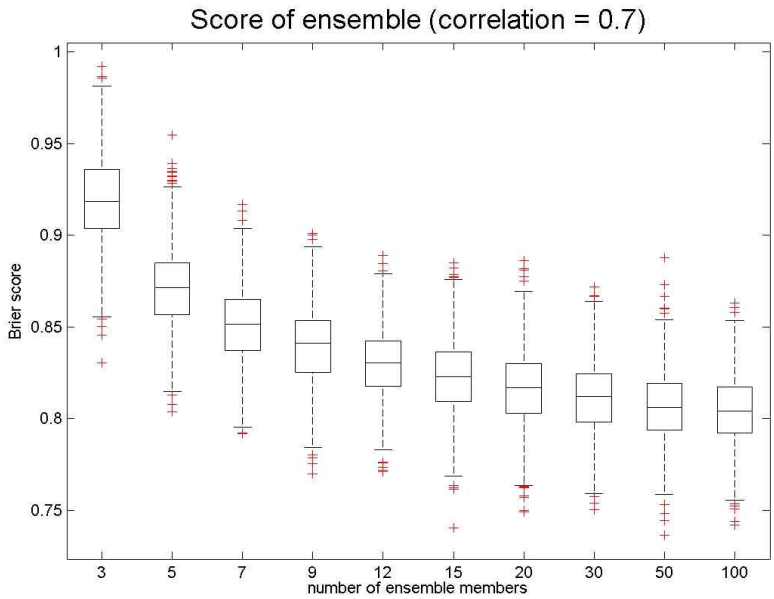


Fig. 4.78 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.3$)

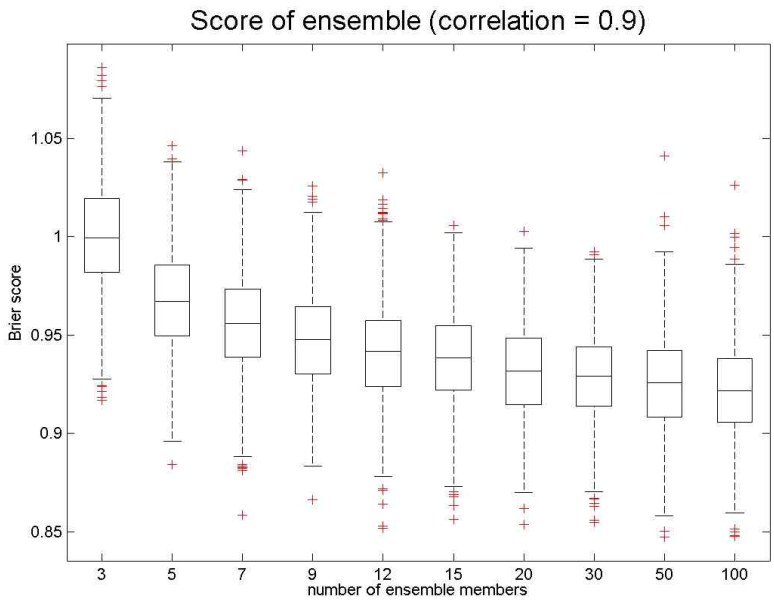


Fig. 4.79 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.3$)

③ $\rho_a = 0.5$

그림 4.80부터 그림 4.85은 ρ_a가 0.5인 경우의 BS 결과를 boxplot으로 나타낸 것이다. x축은 앙상블 내 시나리오 수를 나타내며, y축은 BS를 나타낸다. 그리고 시나리오 간 상관성(ρ_e)별로 결과를 분류하였다. ρ_a가 0.5인 경우도 ρ_a가 0.1과 0.3인 경우와 비슷한 양상을 보인다. 그러므로 이 경우도 앙상블 예측 시스템이 비교적 부정확한 예측을 수행하는 상황을 모의한 것이다. 그러나 ρ_a가 0.5로 향상되어 모든 경우의 ρ_e와 시나리오 수에서 BS가 감소되어 나타나는 것을 확인할 수 있다.

시나리오 간 상관성(ρ_e)이 증가할수록 전체적으로 BS가 커지는 것을 볼 수 있다. 이 경우도 앞선 경우와 마찬가지로 비교적 부정확한 예측을 수행하고 있으므로 시나리오 간 상관성이 앙상블 예측 시스템의 정보 손실을 야기하여 오차를 증가시키는 것으로 볼 수 있다. 또한 시나리오 간 상관성 증가가 시나리오 수 증가로 인한 예측 향상효과를 감소시킨다. 또한 상관성이 증가함에 따라 boxplot의 크기가 커지는 것이 나타난다. 그리고 시나리오 수 사이의 boxplot 크기 차이가 감소한다. 시나리오 간 상관성과 시나리오 수 변화에 따른 앙상블 예측 시스템의 거동들이 앞서 언급한 ρ_a가 0.1과 0.3인 경우와 동일하다.

앙상블을 SE로 평가하여 평균 앙상블의 예측 정확도를 검토한 경우에는 ρ_a가 0.5인 경우에서 다른 결과 양상을 보였다. 시나리오 수와 시나리오 간 상관성이 달라져도 SE의 결과의 변화가 나타나지 않았다. 그러나 BS로 평가하여 확률예측으로 앙상블을 평가하는 경우 ρ_a가 0.1, 0.3과 같은 양상으로 나타났다. 시나리오 간 상관성과 시나리오 수가 미치는 영향이 확정적 예측과 확률예측의 차이에 의해 다르게 나타날 수 있음을 의미한다.

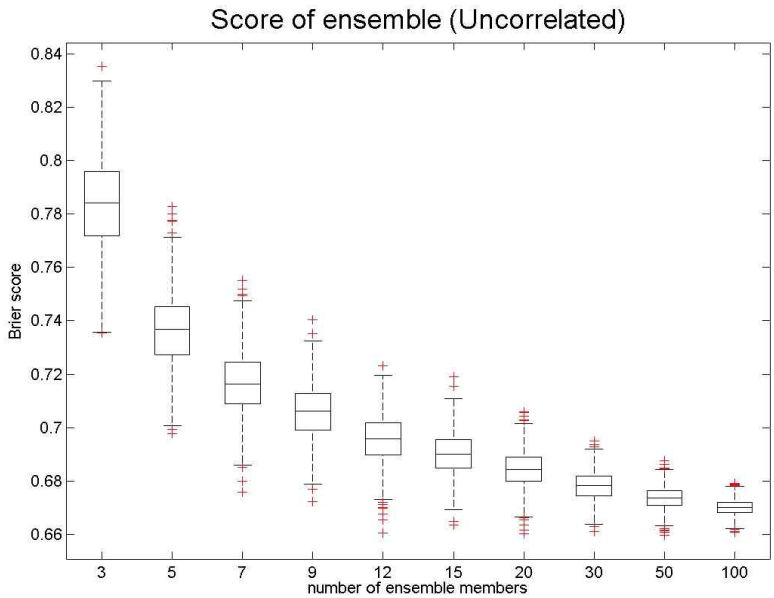


Fig. 4.80 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.5$)

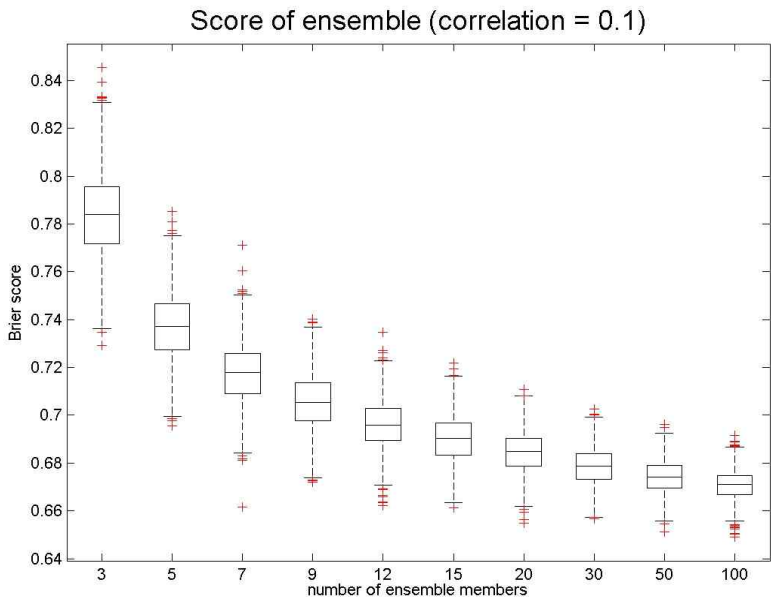


Fig. 4.81 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.5$)

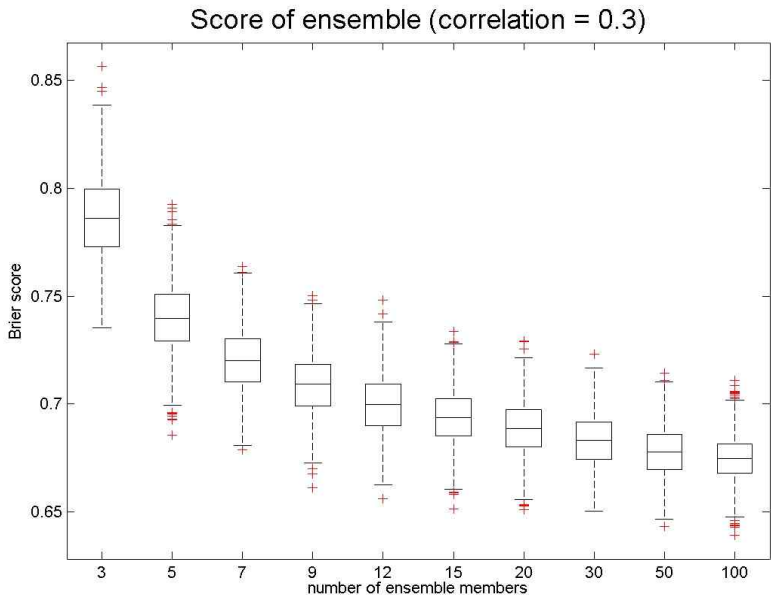


Fig. 4.82 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.5$)

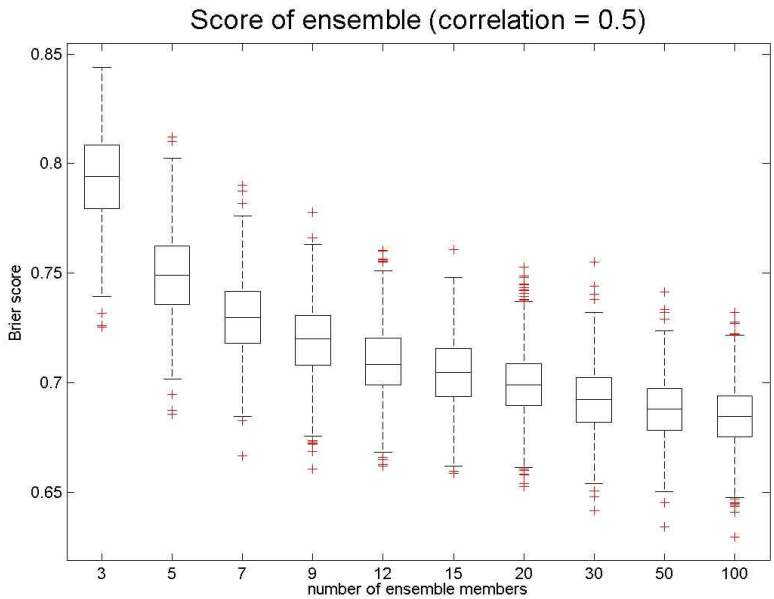


Fig. 4.83 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.5$)

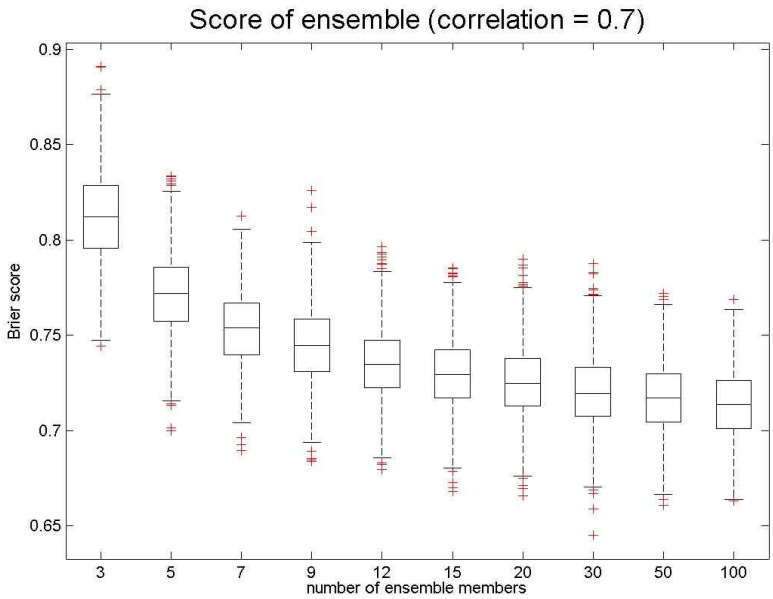


Fig. 4.84 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.5$)

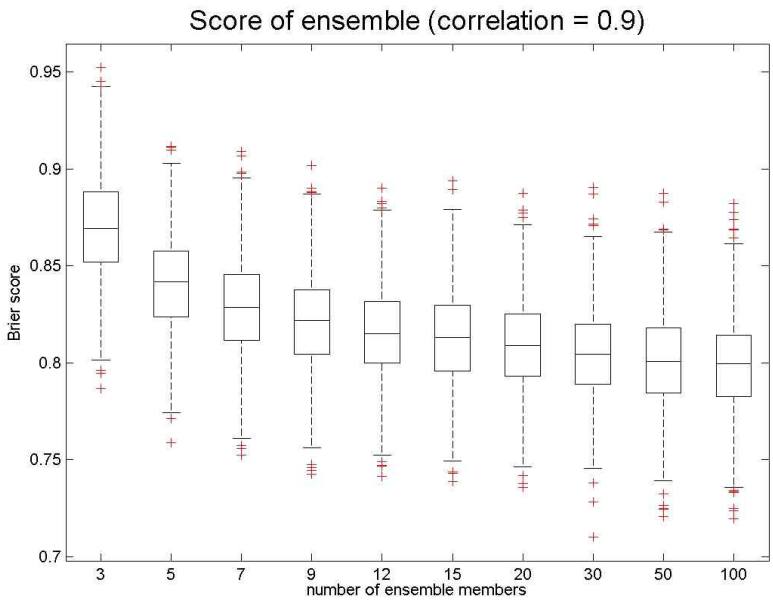


Fig. 4.85 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.5$)

④ $\rho_a = 0.7$

그림 4.86부터 그림 4.91은 ρ_a가 0.7인 경우의 BS 결과를 boxplot으로 나타낸 것이다. x축은 앙상블 내 시나리오 수를 나타내며, y축은 BS를 나타낸다. 그리고 시나리오 간 상관성(ρ_e)별로 결과를 분류하였다. 이 경우에는 앞선 경우와는 조금 다른 양상이 나타난다.

앙상블 시나리오 수가 증가하면 BS가 작아진다. 이는 앞선 ρ_a가 0.1, 0.3, 그리고 0.5인 경우와 일치하는 결과이다. 하지만 ρ_e에 따른 변화는 앞선 경우와 다르게 나타난다. 앞의 경우에서는 ρ_e가 커지면서 오차가 커지는 것을 확인할 수 있었다. 하지만 ρ_a가 0.7인 경우에서는 시나리오 간 상관성(ρ_e)의 크기순서와 BS크기순서의 관계를 찾아볼 수 없다.

가장 BS가 크게 나타난 결과는 ρ_e가 0일 때로 이전 결과와 정반대의 결과가 나타났다. 그리고 그 다음으로 BS가 크게 나온 경우를 차례로 나열하면 0.1, 0.9, 0.3, 0.5, 0.7의 순서로 나타났다. 시나리오 간 상관성과 오차의 연결고리를 찾을 수 없는 결과가 나타났다. ρ_a가 0.7인 경우는 앞의 ρ_a가 0.1, 0.3, 그리고 0.5인 경우에 비하여 높은 예측의 정확도가 보장되는 경우를 의미한다.

확정적 예측 평가기법인 SE로 앙상블을 평가하였을 때 ρ_a가 0.5인 경우에서 시나리오 간 상관성의 변화가 앙상블 예측 정확도에 영향을 미치지 못하는 구간이 나타났다. 확률예측으로서 앙상블을 평가하였을 때 확정적 예측과는 다른 ρ_a에서 이러한 구간이 나타났다. 또한 확정적 예측과는 다르게 ρ_e크기에 무질서한 구간에서도 많은 수의 시나리오를 사용하는 것이 예측 정확도 향상에 도움이 되는 것으로 나타났다.

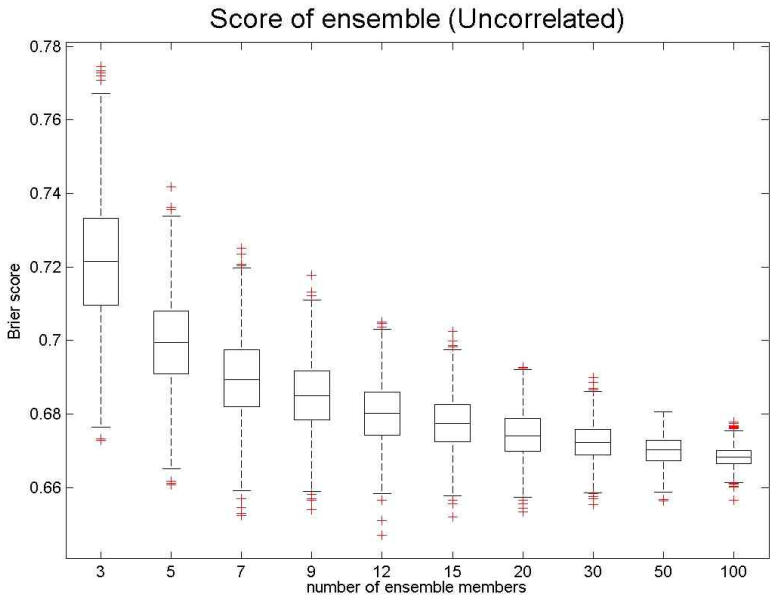


Fig. 4.86 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.7$)

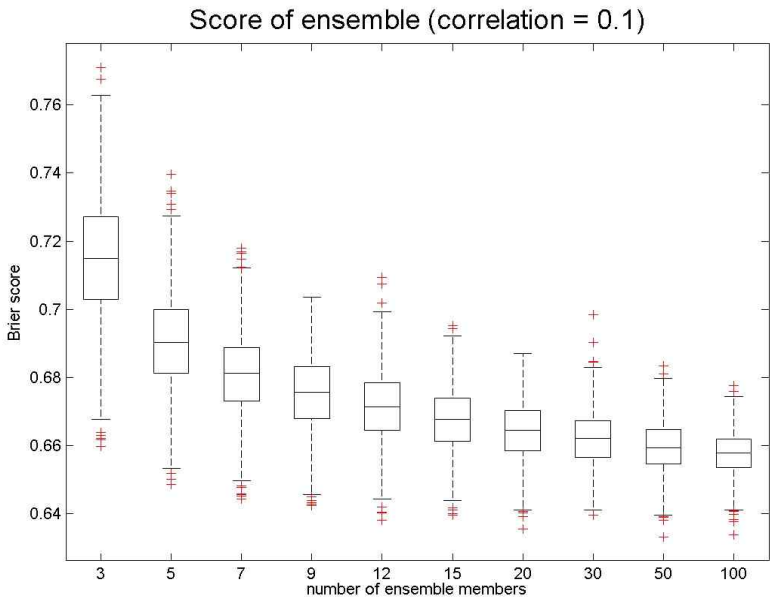


Fig. 4.87 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.7$)

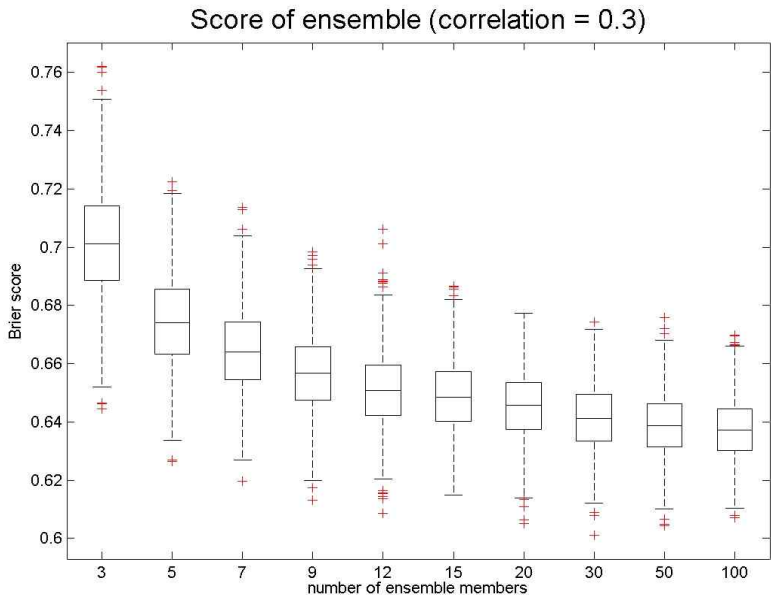


Fig. 4.88 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.7$)

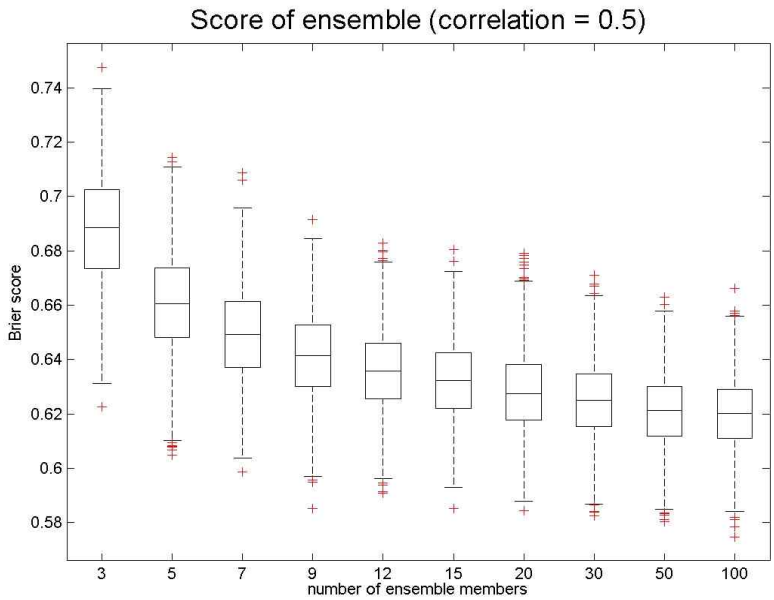


Fig. 4.89 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.7$)

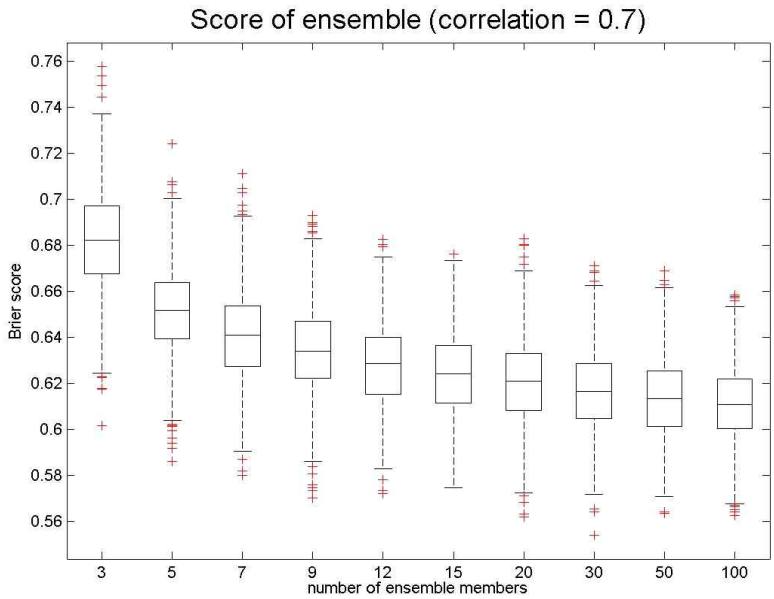


Fig. 4.90 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.7$)

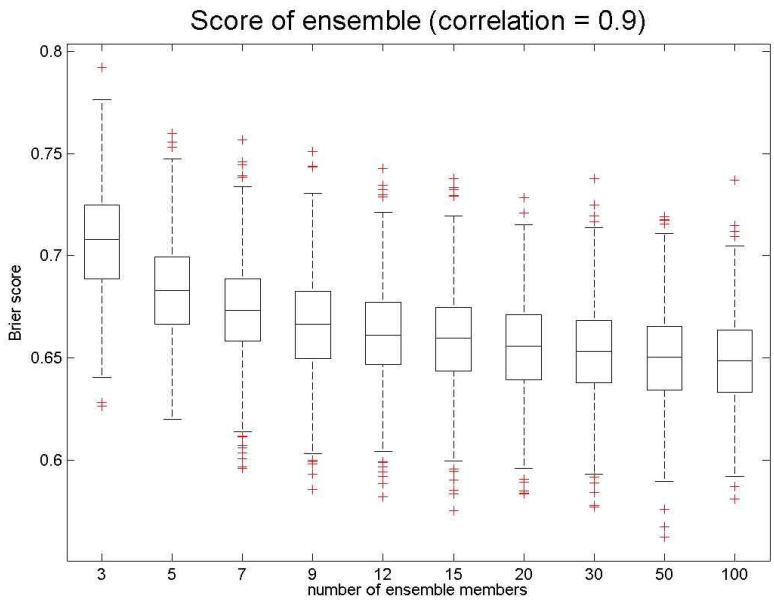


Fig. 4.91 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.7$)

⑤ $\rho_a = 0.9$

그림 4.92부터 그림 4.97은 ρ_a가 0.9인 경우의 BS 결과를 boxplot으로 나타낸 것이다. x축은 앙상블 내 시나리오 수를 나타내며, y축은 BS를 나타낸다. 그리고 시나리오 간 상관성(ρ_e)별로 결과를 분류하였다. ρ_a가 0.9인 경우로 앙상블 예측 시스템 내에 정확한 예측을 수행하는 시나리오가 반드시 하나 이상 존재하는 상황을 모의한 것이다. 앞서 모의한 ρ_a가 0.1, 0.3, 0.5, 그리고 0.7인 경우와는 상당히 다른 양상을 보인다.

ρ_a가 0.9일 때는 비교적 예측 정확도가 낮은 경우인 ρ_a가 0.1, 0.3, 0.5인 경우에 비교하여 시나리오 간 상관성에 따른 BS의 변화가 정반대의 양상이 나타난다. ρ_e가 0에서 0.9로 증가함에 따라 예측의 정확도가 향상되고 있음을 확인할 수 있다. 예측의 정확도가 높은 시나리오가 하나 이상 보장되는 경우 시나리오 간 상관성이 높을수록 더 좋은 예측을 수행하는 것으로 나타났다. 높은 예측의 정확도를 가지는 시나리오 $\boldsymbol{y_1}$과 상관 정도가 높을수록 정확한 예측을 하는 시나리오가 많아지게 된다. 따라서 이와 같이 정확한 예측을 반드시 수행할 수 있는 경우 일반적인 경우와 다르게 시나리오 간 상관성이 높을수록 예측의 정확도가 높아지게 된다.

시나리오 간 상관성 ρ_e가 0, 0.1, 0.3, 그리고 0.5인 경우에는 앙상블 내 시나리오 수가 증가할수록 BS가 커져 예측의 정확도가 저하되는 결과가 나타났다. ρ_e가 0.7, 0.9인 경우에는 이전의 결과들과 마찬가지로 시나리오 수가 증가함에 따라 예측의 정확도가 향상되었다. 앙상블 내에 정확한 예측 시나리오가 최소한 하나가 존재하는 경우 시나리오 수의 영향은 시나리오 간 상관성(ρ_e)에 의해서 예측의 정확도에 다르게 작용할 수 있다. ρ_a가 0.9이고, 시나리오 간 상관성이 낮은 경우에 시나리오 수 증가는 $\boldsymbol{y_1}$에 비하여 상대적으로 부정확한 시나리오의 증가를 의미하므로 BS가 점차 커지게 된다.

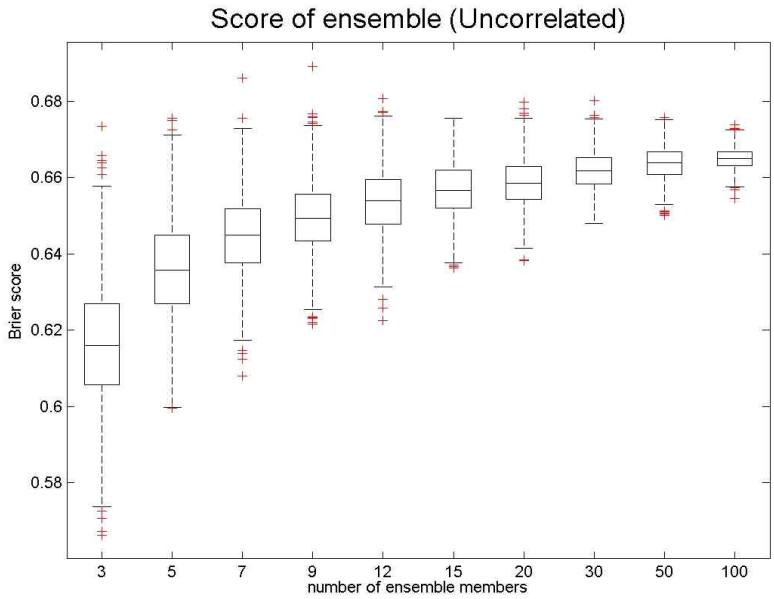


Fig. 4.92 Brier score of correlated ensemble ($\rho_e = 0,\ \rho_a = 0.9$)

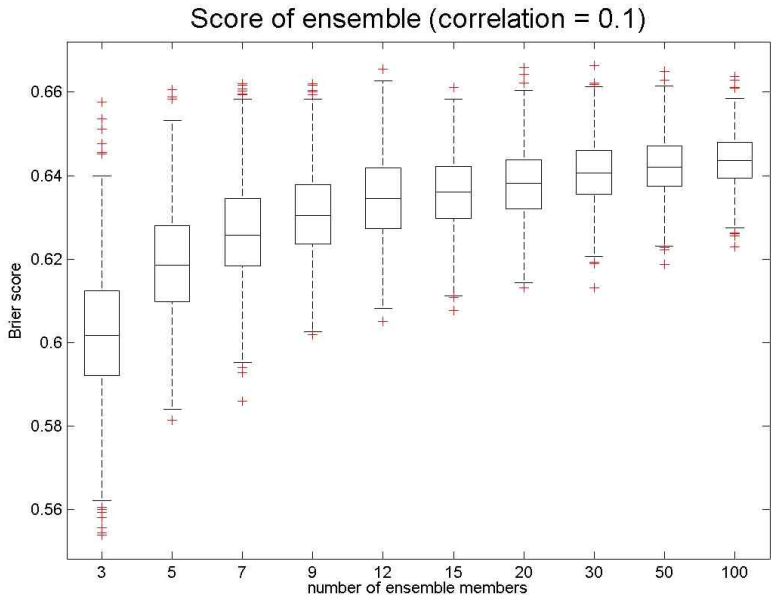


Fig. 4.93 Brier score of correlated ensemble ($\rho_e = 0.1,\ \rho_a = 0.9$)

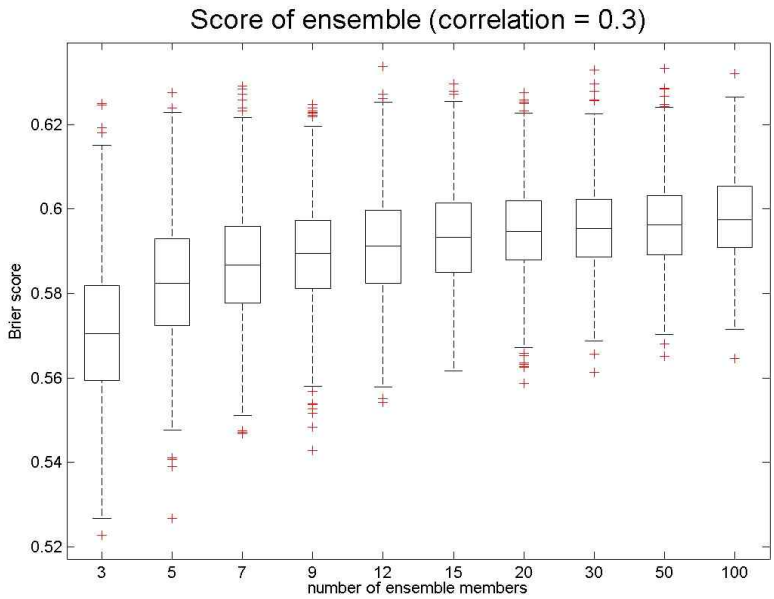


Fig. 4.94 Brier score of correlated ensemble ($\rho_e = 0.3,\ \rho_a = 0.9$)

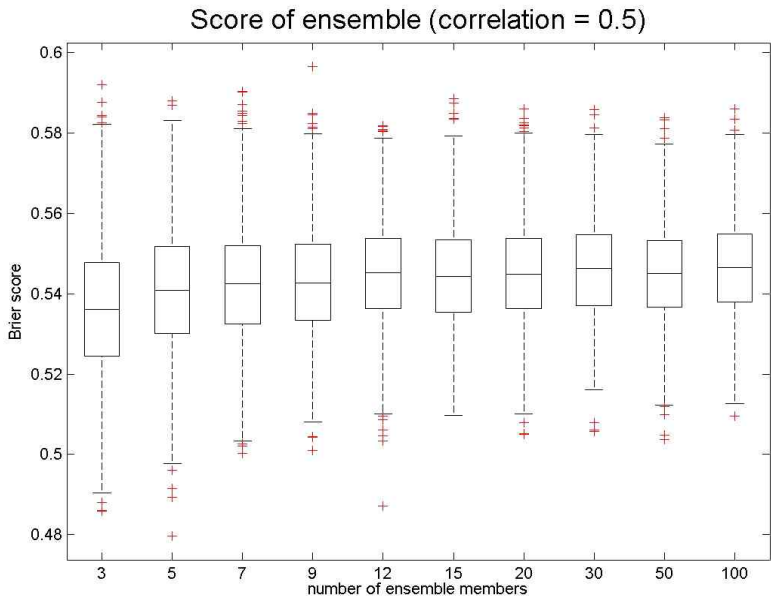


Fig. 4.95 Brier score of correlated ensemble ($\rho_e = 0.5,\ \rho_a = 0.9$)

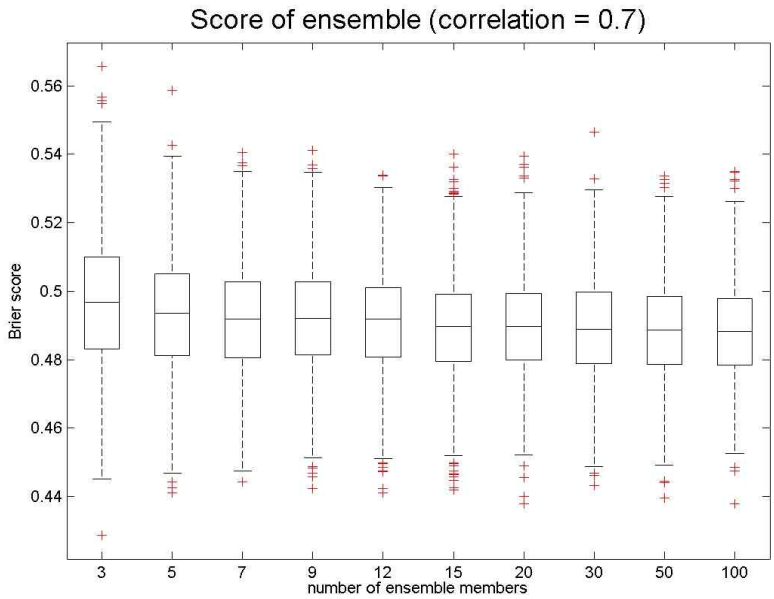


Fig. 4.96 Brier score of correlated ensemble ($\rho_e = 0.7,\ \rho_a = 0.9$)

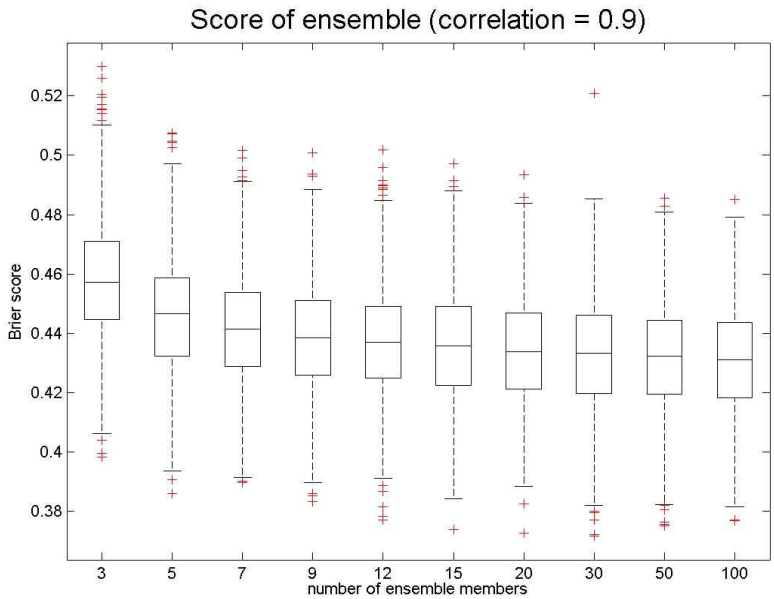


Fig. 4.92 Brier score of correlated ensemble ($\rho_e = 0.9,\ \rho_a = 0.9$)

4.3.3 앙상블 시나리오 간 상관성 효과분석 결과

앙상블 시나리오 간 상관성이 예측 시스템에 미치는 영향을 분석하기 위하여 시나리오 간 상관성(ρ_e)을 0, 0.1, 0.3, 0.5, 0.7, 그리고 0.9로 조정하여 앙상블을 생성하였다. 그리고 앙상블 내 시나리오 $\boldsymbol{y_1}$과 관측시나리오의 상관계수인 ρ_a를 0.1, 0.3, 0.5, 0.7, 그리고 0.9로 정하여 관측값을 생성하여 효과를 분석하였다.

생성된 상관 앙상블을 평가하는 지표로는 확정적 예측 평가방법과 확률적 예측 평가방법을 사용하였다. 확정적 예측 평가 방법은 앙상블 예측에서 생성된 단일 예측을 평가하는 방법이다. 확정적 예측 평가지표로는 D^2와 SE를 사용하였다. D^2는 상관 앙상블이 얼마나 퍼져있는가를 평가하는 지표로 관측시나리오와는 무관하게 앙상블 자체의 특성을 보여준다. SE는 평균 앙상블이 가지는 오차를 평가하는 지표이다. 확률적 예측지표로는 BS가 사용되었다. 본 절에서는 시나리오 간 상관성이 앙상블 예측 시스템에 미치는 영향을 정리하였다.

확정적 예측 평가지표로 평가를 할 경우, 전체 시스템의 오차는 SE와 D^2의 합으로 이루어져 있으며 D^2가 증가하면 평균 앙상블 예측의 오차가 감소할 수 있음을 의미한다. 일반적으로 앙상블에서 D^2가 클수록 좋은 예측 시스템이라고 할 수 있다.

D^2로 평가하였을 경우 결과는 다음과 같다. ρ_e가 0에서 0.9까지 커짐에 따라서 D^2가 작아지는 결과를 보여주었다(그림 4.12). 상관성이 커질수록 앙상블 시나리오 간 거리가 가까워졌다. 또한 ρ_e가 커질수록 boxplot의 크기가 작아져, 시나리오 간 상관성이 커질수록 가능한 D^2값의 범위가 작아진다는 것을 보여주었다(그림 4.13~그림 4.22). 따라서 상관성의 증가는 표현할 수 있는 예측 범위를 감소시킨다는 것을 알 수 있다. 시나리오 수가 작은 구간에서 수가 증가할 때 D^2가 증가하는 정도가 커지며, 일정 수 이상에서는 시나리오

수 증가가 앙상블 시스템이 표현할 수 있는 예측 범위 증가에 크게 영향을 미치지 못함을 보여준다(그림 4.23~그림 4.27).

SE를 이용하여 앙상블을 평가한 경우에 ρ_a별 결과와 전체 결과로 나누어 분석하였다. ρ_a별 결과는 $\boldsymbol{y_1}$의 정확도를 알 수 있어 앙상블 예측의 정확도를 어느 정도 알 수 있는 경우를 모의한 것이고, 앙상블 예측의 정확도를 알 수 없는 경우를 모든 경우의 ρ_a 결과를 한꺼번에 분석하여 전체 분석 결과로 나타내었다.

SE 전체 분석 결과는 다음과 같다. 시나리오 수가 증가할수록 boxplot의 크기가 작아지는데, 오차의 분산이 시나리오 수가 증가할수록 작아지는 것을 알 수 있다. 또한 결과의 중앙값이 시나리오 수에 관계없이 거의 1 값에 고정되어있다. 따라서 평균 앙상블 시나리오를 사용했을 시에 시나리오 수가 많을수록 안정적으로 더 좋은 예측을 할 수 있다는 것을 알 수 있다. 시나리오 수가 적을수록 불확실성이 크다고 할 수 있다. 시나리오 간 상관성(ρ_e)이 커질수록 boxplot의 크기 커진다. 이는 시나리오 간 상관계수의 증가로 앙상블 예측 시스템이 가질 수 있는 정보의 손실로 인해 불확실성이 커진 것을 의미한다. 또한 시나리오 간 상관성이 커질수록 시나리오 수 증가의 효과가 점점 약해지는 것이 나타난다.

ρ_a별 SE 분석 결과는 다음과 같다. ρ_a가 0.1에서 0.9까지 증가하면서 전체적으로 오차가 줄어드는 것이 나타난다. ρ_a가 0.1과 0.3인 경우에는 시나리오 간 상관성이 작을수록 그리고 시나리오 수가 많을수록 오차가 작아지는 것을 확인할 수 있다. ρ_a가 0.5인 경우에서는 시나리오 간 상관성과 시나리오 수가 변하여도 SE 결과가 변하지 않는 현상이 나타난다. ρ_a가 0.7과 0.9인 경우에는 시나리오 간 상관성이 클수록 그리고 시나리오 수가 적을수록 오차가 작아지는 것을 확인할 수 있다.

BS를 이용하여 상관 앙상블을 평가한 경우에도 ρ_a별 결과와 전체 결과로 나누어 분석하였다. BS의 전체 분석 결과는 다음과 같다.

시나리오 간 상관성(ρ_e)이 증가하면 전반적으로 BS 평균이 커지며, boxplot의 크기가 커져 분산이 증가한다(그림 4.61～그림 4.65). 시나리오 간 상관성 증가로 인하여 정보의 손실이 발생하고, 앙상블 시스템이 가능한 예측의 폭이 줄어들게 된다. 이로 인하여 앙상블이 더 극단적인 예측을 수행하게 됨을 알 수 있다. 많은 수의 시나리오를 사용함으로써 이러한 극단적인 예측을 방지하게 된다. 하지만 ρ_e가 증가함으로써 시나리오 수 증가로 인한 효과가 감소하여 boxplot 간 크기의 차이가 거의 나타나지 않으며(그림 4.65), 시나리오 수 대비 boxplot의 평균의 차이가 점차 감소한다.

ρ_a별 BS 분석 결과는 다음과 같다. 분석 결과를 요약하기 위하여 각 boxplot의 평균을 구하여 나타내었다(그림 4.93～그림 4.97). ρ_a가 0.1, 0.3, 그리고 0.5인 경우에는 시나리오 간 상관성(ρ_e)이 작을수록 BS가 작게 나타난다. 그리고 시나리오 수가 많을수록 예측이 향상되는 것을 확인할 수 있다. ρ_a가 0.7인 경우에는 ρ_e의 크기와 예측의 정확도에 관계성이 나타나지 않는다. ρ_a가 0.9인 경우에는 앞의 경우와는 반대로 ρ_e가 클수록 BS가 작게 나타난다. ρ_e가 0.7과 0.9인 경우에만 시나리오 수의 증가가 예측의 정확도 향상에 도움이 되었으며, 나머지 ρ_e의 경우에서는 시나리오 수가 증가하면 BS가 커지는 결과가 나타났다.

앙상블의 예측의 정확도를 예상할 수 없는 경우에 시나리오 간 상관성이 작을수록, 시나리오 수가 많을수록 더 좋은 예측을 수행할 가능성이 높아진다. 그리고 시나리오 간 상관성이 높을수록 많은 수의 시나리오를 사용하는 것이 무의미해지는 결과가 나타난다. ρ_a별로 분석을 하면 세 가지 상황을 생각할 수 있다. 비교적 정확도가 낮은 예측이 예상될 경우에는 시나리오 수가 많고 상관성이 작을수록 좋은 결과가 나타난다. 비교적 높은 정확도의 예측을 수행할 수 있을 때는 상관성이 많을수록 더 좋은 결과를 얻을 수 있다. 또한 시나리오 수의 증가가 오히려 예측의 정확도를 낮추는 결과를 보이

기도 한다. 즉 높은 정확도의 예측이 보장되는 경우에는 앙상블 예측이 오히려 하나의 예측 시나리오만을 제공하는 예측 방법보다 더 낮은 예측 수행력을 보일 수 있다는 것을 알 수 있다. 세 번째 상황은 양쪽 모두 아닌 경우이다. 높은 예측을 수행할 수 있거나 불확실성이 높아 낮은 예측밖에 수행할 수 없는 경우 외에 그 중간 정도 되는 예측을 수행하는 상황에서는 시나리오 간 상관성이 앙상블 예측의 정확도에 영향을 미치지 못하는 결과가 나타난다.

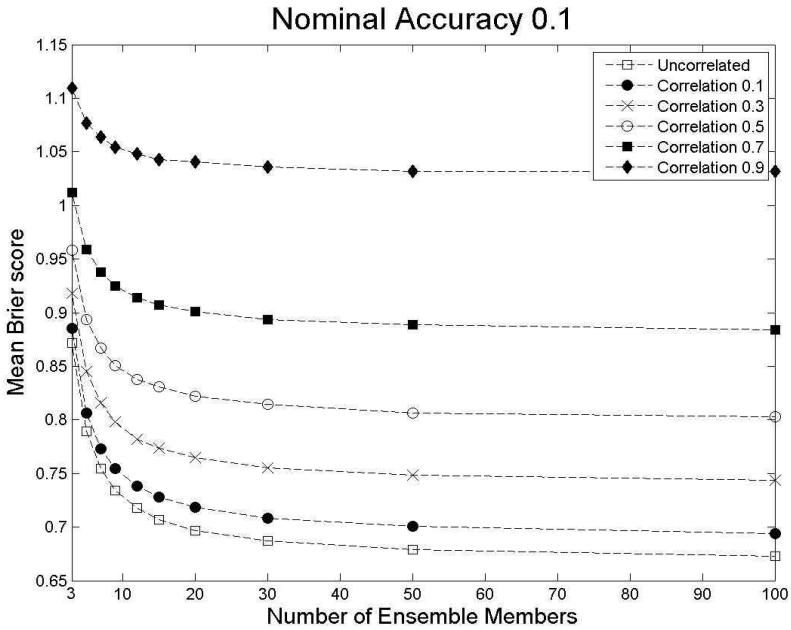


Fig. 4.93 Summary of Brier score results ($\rho_a = 0.1$)

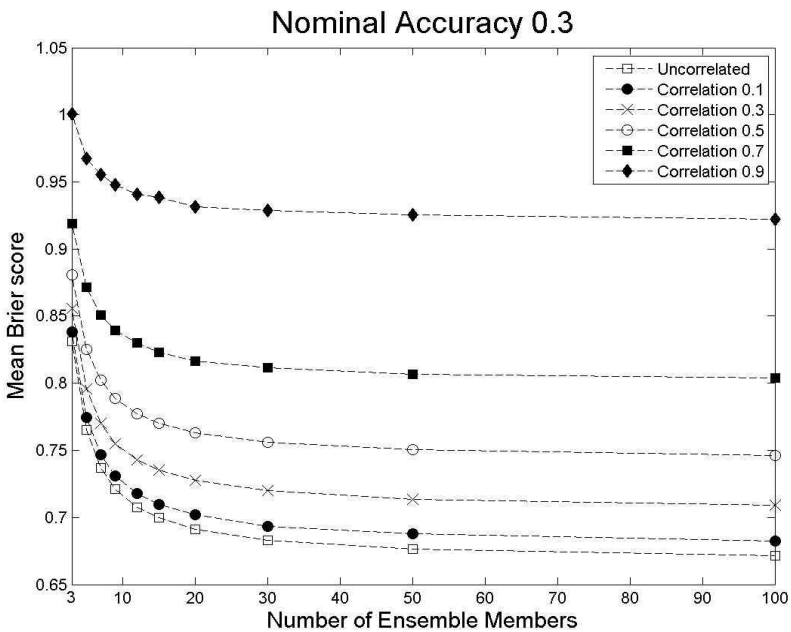


Fig. 4.94 Summary of Brier score results ($\rho_a = 0.3$)

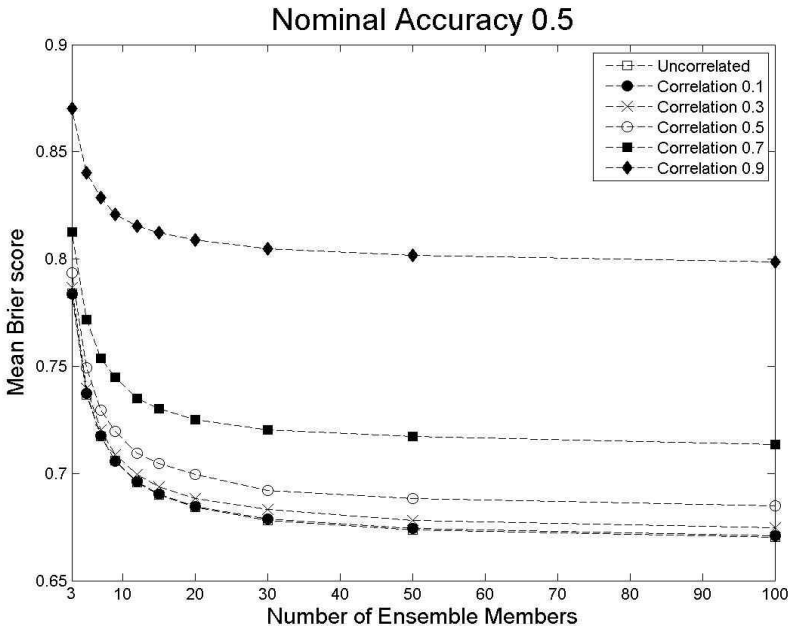


Fig. 4.95 Summary of Brier score results ($\rho_a = 0.5$)

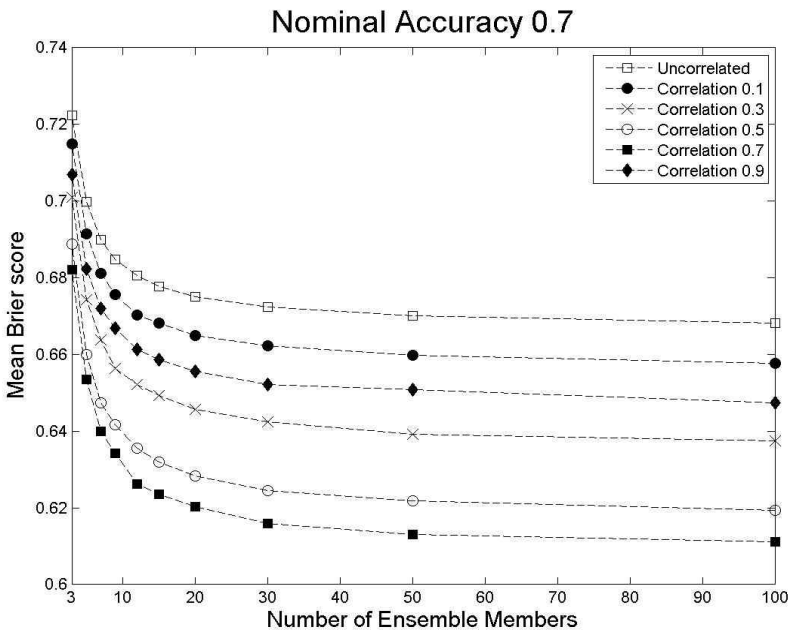


Fig. 4.96 Summary of Brier score results ($\rho_a = 0.7$)

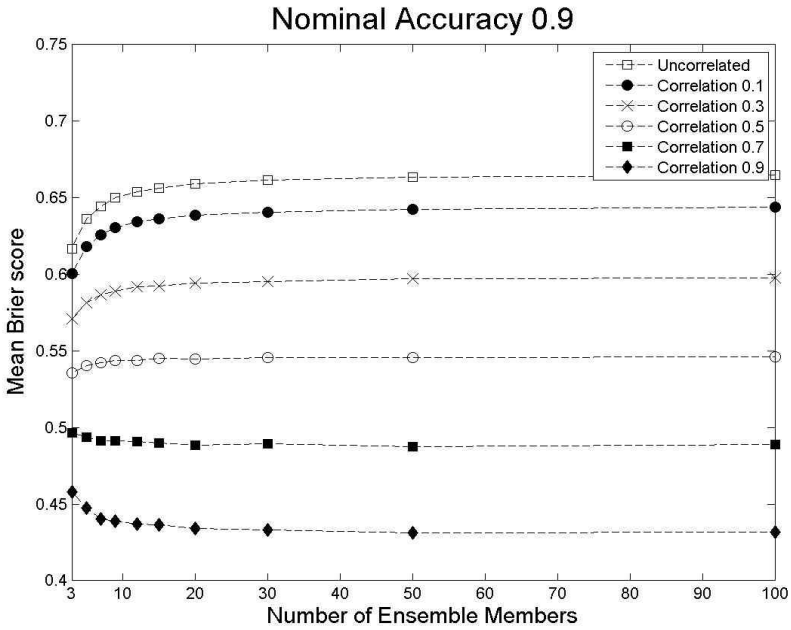


Fig. 4.97 Summary of Brier score results ($\rho_a = 0.9$)

4.4 유효 앙상블 시나리오 수 추정

4.4.1 90% 향상 구간 추정

BS가 시나리오 수의 증가에 따라 향상되는 정도를 구간별로 비교하여 분석하였다. 구간별 예측의 정확도 향상을 비교하기 위하여 각 시나리오 간 상관성(ρ_e)과 시나리오 $\boldsymbol{y_1}$의 정확도 ρ_a별로 BS의 최대값과 최소값의 차이의 90%가 되는 지점을 구하여 유효 앙상블 시나리오 수를 추정하였다. 예를 들어 앙상블 시나리오 수가 증가할수록 앙상블 예측의 정확도가 향상되는 경우에 멤버 수가 3인 앙상블과 100인 앙상블의 스코어 차이의 90%를 계산하였다. 그리고 그 스코어에 해당하는 앙상블 수를 추정하였다.

그림 4.98부터 그림 4.101은 유효 앙상블 시나리오 수를 추정하기 위하여 사용한 BS결과 그래프이다. ρ_a별로 결과를 나타내었으며, 각 결과는 다시 시나리오 간 상관성별로 구분된다. x축은 앙상블 내 시나리오 수를 나타내며 y축은 1000회씩 모의된 결과를 평균하여 나타낸 것이다. ρ_a는 0.1, 0.3, 0.5, 그리고 0.7인 경우에 대하여 유효 앙상블 수를 추정하였다. 이 경우에 시나리오 수가 증가할수록 BS가 작아져 앙상블 예측의 정확도가 증가하기 때문이다. ρ_a가 0.9인 경우에는 대부분 시나리오 수가 증가하면 예측의 정확도가 낮아지기 때문에 유효 앙상블 수가 존재할 수 없다.

시나리오 수가 3인 경우의 BS에서 시나리오 수가 100인 경우의 BS까지 향상된 점수의 90%에 해당하는 점수를 찾은 뒤 그 점수에 해당되는 시나리오 수를 추정하였다. 추정된 지점은 붉은색 십자로 표시하였다. 모든 ρ_a와 ρ_e에 대한 경우의 그래프에서 90% 향상 지점이 시나리오 수 20개에서 30개 사이에 존재하며, 대부분의 경우 25개로 추정되었다(표 4.1).

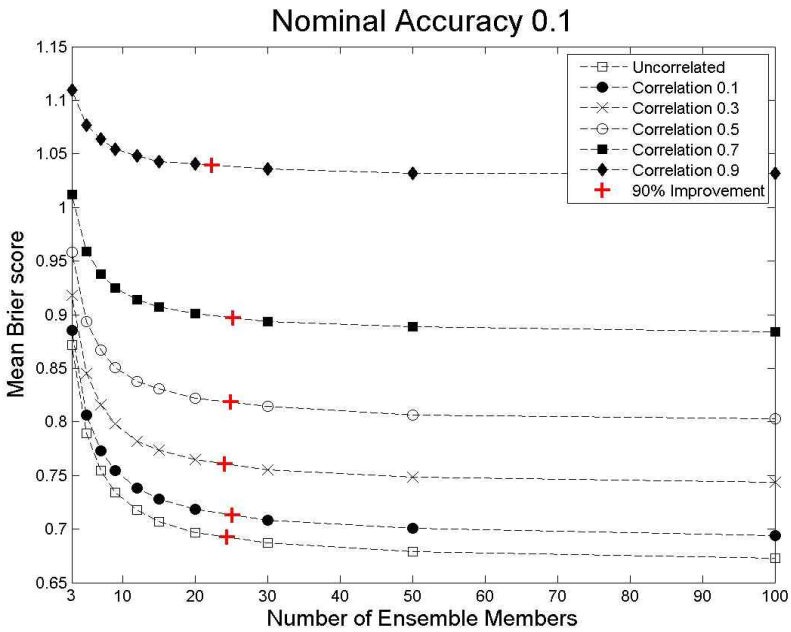


Fig. 4.98 The effective number of scenarios ($\rho_a = 0.1$)

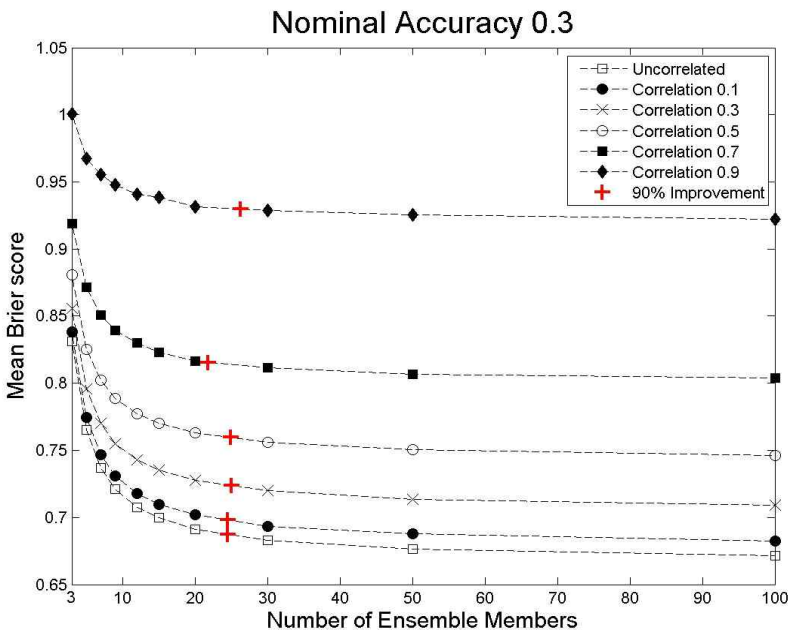


Fig. 4.99 The effective number of scenarios ($\rho_a = 0.3$)

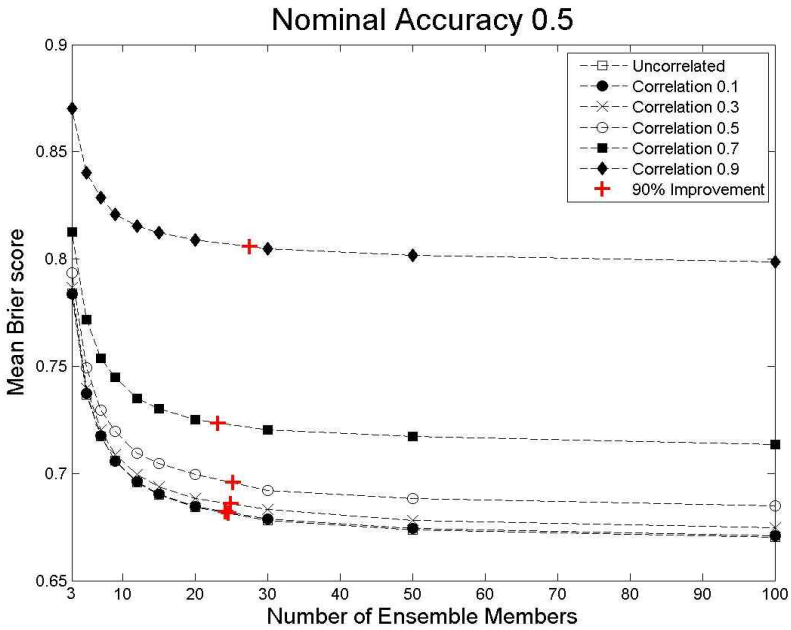


Fig. 4.100 The effective number of scenarios ($\rho_a = 0.5$)

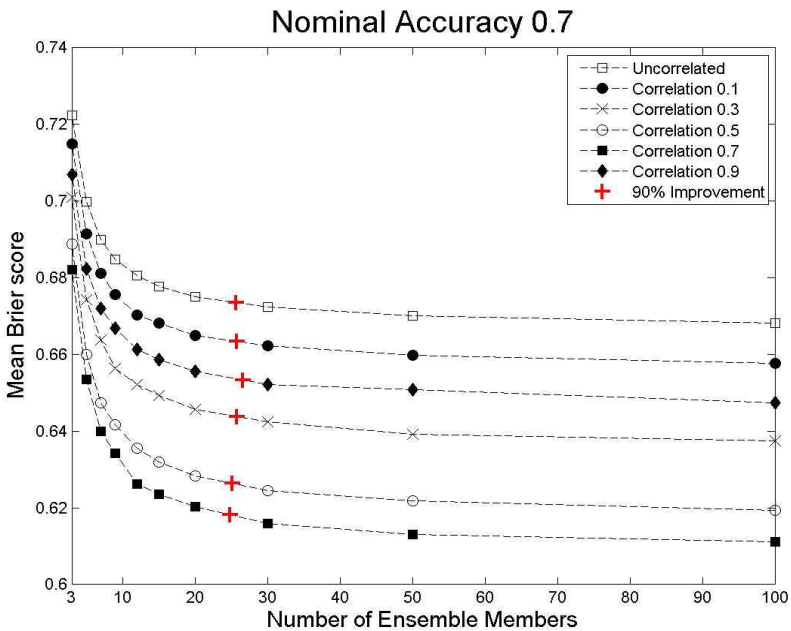


Fig. 4.101 The effective number of scenarios ($\rho_a = 0.9$)

Table 4.1 The effective number of ensemble scenarios

Correlation between ensemble scenarios		0	0.1	0.3	0.5	0.7	0.9
$\rho_a = 0.1$	Number of scenarios	24.4	25.1	24.1	24.9	25.3	22.3
	Brier score (90%)	0.69	0.71	0.76	0.82	0.90	1.04
$\rho_a = 0.3$	Number of scenarios	24.5	24.5	25.1	24.9	21.8	26.2
	Brier score (90%)	0.69	0.70	0.72	0.76	0.82	0.93
$\rho_a = 0.5$	Number of scenarios	24.6	24.2	25.0	25.2	23.2	27.5
	Brier score (90%)	0.68	0.68	0.69	0.70	0.72	0.81
$\rho_a = 0.7$	Number of scenarios	25.7	25.8	25.7	25.1	24.8	26.6
	Brier score (90%)	0.67	0.66	0.64	0.63	0.62	0.65

4.4.2 구간별 향상률 비교

BS가 향상되는 정도를 그래프의 기울기를 구하여 판단하였다. 각 ρ_e와 ρ_a별로 평균 BS의 시나리오 수 증가 대비 변화율을 구하였다. 이때의 기울기가 시나리오 수가 증가하는 정도에 앙상블 예측의 정확도가 향사되는 정도를 가장 정확하게 보여주는 지표라고 볼 수 있다. 그러므로 일정이상의 기울기, 즉 향상 정도를 가지지 못하는 지점에서 유효 앙상블 시나리오 수라고 판단하였다. 대부분의 경우에서 3개에서 5개로 시나리오 수가 늘어났을 때 가장 스코어가 크게 향상되었다. 이때의 기울기의 5%가 되는 지점을 유효 앙상블 멤버 수 판단의 근거로 사용하였다. 그 이하의 기울기를 보이는 구간은 시나리오 대비 예측 정확도 향상 정도가 비효율적이라고 판단하였다.

ρ_a와 ρ_e별로 각각의 평균 BS 결과 그래프의 기울기는 표 4.2에 나타나있다. ρ_a가 0.9인 경우에는 앙상블을 사용하는 것이 비효율적이라는 결과가 나타났으므로 ρ_a가 0.1, 0.3, 0.5, 그리고 0.7인 경우에 대하여 기울기를 구하였다. 모든 경우에 시나리오 수가 증가할수록 그래프의 기울기가 작아지는 것을 알 수 있다. 시나리오 수가 증가할수록 증가 대비 예측 정확도 향상 정도가 작아지는 것을 확인할 수 있다. 또한 ρ_a가 작을수록 기울기가 전반적으로 큰 결과가 나타난다. 이것은 낮은 정확도의 예측을 수행할 수밖에 없는 경우에 앙상블 시나리오 수를 증가시키는 것이 높은 경우에 비하여 효율이 좋다는 것을 의미한다.

표 4.2의 제일 오른쪽 열은 첫 번째 구간 기울기의 5%를 나타내며, 이와 가장 근사한 구간의 기울기를 밑줄로 표시하였다. 대부분의 경우에서 15개와 20개 사이의 시나리오 수 구간에서 5% 기울기가 나타났다. 이후의 기울기는 첫 구간의 5%가 되지 못하며, 효율이 아주 작다는 것을 보여준다. 기울기로 판단하였을 때 20개 정도의 시나리오를 사용하는 것이 적절하다고 사료된다.

Table 4.5 The slope of Brier score curve between each interval

Accuracy	Correlation	|Slope| Interval of Scenario number									5% Slope
		3~5	5~7	7~9	9~12	12~15	15~20	20~30	30~50	50~100	
0.1	0	0.0411	0.0175	0.0099	0.0057	0.0035	0.0020	0.0010	0.0004	0.0001	0.0021
	0.1	0.0395	0.0168	0.0091	0.0054	0.0035	0.0019	0.0010	0.0004	0.0001	0.0020
	0.3	0.0363	0.0147	0.0088	0.0054	0.0029	0.0017	0.0010	0.0004	0.0001	0.0018
	0.5	0.0326	0.0131	0.0082	0.0043	0.0023	0.0018	0.0007	0.0004	0.0001	0.0016
	0.7	0.0264	0.0106	0.0065	0.0038	0.0022	0.0012	0.0008	0.0003	0.0001	0.0013
	0.9	0.0164	0.0066	0.0046	0.0020	0.0018	0.0005	0.0004	0.0002	0.0000	0.0008
0.3	0	0.0331	0.0142	0.0079	0.0045	0.0026	0.0017	0.0008	0.0003	0.0001	0.0017
	0.1	0.0318	0.0140	0.0079	0.0043	0.0028	0.0015	0.0008	0.0003	0.0001	0.0016
	0.3	0.0300	0.0127	0.0076	0.0041	0.0025	0.0015	0.0008	0.0003	0.0001	0.0015
	0.5	0.0276	0.0116	0.0067	0.0039	0.0024	0.0014	0.0007	0.0003	0.0001	0.0014
	0.7	0.0239	0.0101	0.0057	0.0032	0.0023	0.0013	0.0005	0.0002	0.0000	0.0012
	0.9	0.0167	0.0060	0.0038	0.0023	0.0008	0.0014	0.0003	0.0002	0.0001	0.0008
0.5	0	0.0236	0.0098	0.0056	0.0034	0.0019	0.0012	0.0006	0.0002	0.0001	0.0012
	0.1	0.0232	0.0098	0.0060	0.0032	0.0020	0.0011	0.0006	0.0002	0.0001	0.0012
	0.3	0.0234	0.0097	0.0058	0.0031	0.0020	0.0010	0.0005	0.0003	0.0001	0.0012
	0.5	0.0221	0.0099	0.0049	0.0035	0.0016	0.0010	0.0007	0.0002	0.0001	0.0011
	0.7	0.0204	0.0090	0.0045	0.0032	0.0017	0.0010	0.0005	0.0002	0.0001	0.0010
	0.9	0.0149	0.0059	0.0038	0.0019	0.0009	0.0007	0.0004	0.0002	0.0001	0.0007
0.7	0	0.0112	0.0050	0.0026	0.0014	0.0010	0.0005	0.0003	0.0001	0.0000	0.0006
	0.1	0.0118	0.0051	0.0028	0.0018	0.0007	0.0007	0.0003	0.0001	0.0000	0.0006
	0.3	0.0133	0.0053	0.0037	0.0014	0.0010	0.0007	0.0003	0.0002	0.0000	0.0007
	0.5	0.0143	0.0063	0.0029	0.0020	0.0012	0.0007	0.0004	0.0001	0.0000	0.0007
	0.7	0.0143	0.0068	0.0029	0.0027	0.0009	0.0007	0.0004	0.0002	0.0000	0.0007
	0.9	0.0123	0.0052	0.0026	0.0018	0.0009	0.0006	0.0003	0.0001	0.0001	0.0006

제 5 장 결론 및 향후 연구

5.1 요약 및 결론

본 연구는 앙상블 예측을 수행할 시에 앙상블 시나리오 간 상관성이 예측의 정확도에 미치는 영향을 분석하였다. 앙상블 예측은 발생 가능한 여러 예측을 사용하는 방법이므로 시나리오 간 상관성은 정보의 손실을 의미한다. 또한 대표 시나리오($\boldsymbol{y_1}$)과 관측시나리오 사이의 상관관계(ρ_a)를 조절하여 발생 가능한 상황을 설정하여 모의를 진행하였다. 그리고 각 상황별로 시나리오 수 증가 대비 예측의 정확도 향상 정도를 분석하여 유효 앙상블 시나리오 수를 찾고자 하였다.

생성된 상관 앙상블을 평가하는 지표로 확정적 예측 평가방법과 확률적 예측 평가방법을 사용하였다. 확정적 예측 평가지표로는 D^2와 SE를 사용하였다. D^2는 상관 앙상블이 얼마나 퍼져있는가를 평가하는 지표로 관측시나리오와는 무관하게 앙상블 자체의 특성을 보여준다. SE는 평균 앙상블이 가지는 오차를 평가하는 지표이다. D^2와 SE의 합은 평균제곱오차(MSE)가 되므로 같은 MSE의 경우에는 D^2가 클수록 평균 앙상블 예측의 정확도가 증가한다. 확률적 예측지표로는 BS가 사용되었다. BS는 하프 브라이어 스코어가 아닌 다중범주 확률예측 평가를 위하여 본래의 브라이어 스코어를 사용하였다.

표준정규분포 사용의 타당성을 검토하기 위하여 실제 유량의 변동계수(C_v)를 사용하여 분석하였다. C_v는 0.1, 0.5, 1을 사용하여 SE로 앙상블의 거동을 확인하였다. 결과적으로 변동계수에 변화가 있어도 SE 결과 그래프 개형에 변화가 없고 지표의 크기만 다르게 되는 것으로 나타났다. 따라서 표준정규분포를 사용하는 것에 문제

가 없음을 확인하였다.

결과는 ρ_a별 분석 결과와 전체 분석 결과로 나눌 수 있다. ρ_a별 분석 결과는 대표 시나리오($\boldsymbol{y_1}$)과 관측시나리오 사이의 상관관계를 알 수 있으므로 앙상블 예측의 정확도를 예상할 수 있는 상황을 나타낸다. ρ_a가 높은 경우에는 높은 정확도의 예측을 수행할 수 있는 상황이며, ρ_a가 낮은 경우는 불확실성이 높아 낮은 정확도의 예측을 수행할 수밖에 없는 상황이다. 전체 분석 결과는 ρ_a별 결과를 모두 합친 것으로 앙상블 예측의 정확도를 예상할 수 없는 경우이며, 시나리오 간 상관성만을 고려한 일반론적 결과라고 볼 수 있다.

시나리오 간 상관성만이 고려된 전체 결과는 다음과 같다. ρ_e가 0에서 0.9까지 커짐에 따라서 D^2가 감소하였으며, boxplot의 크기가 작아졌다. 따라서 시나리오 간 상관성이 증가할수록 앙상블 시나리오 간 거리가 가까워지며 표현할 수 있는 예측의 범위를 감소시킨다는 것을 알 수 있다. 또한 시나리오 수가 증가할수록 D^2가 커지는 결과가 나타났다. 그리고 그 증가율은 시나리오 수가 커질수록 감소하였다. 이는 시나리오 수가 많을수록 예측의 범위가 증가하지만 그 효과는 시나리오 수 증가와 함께 점차 감소한다는 것을 의미한다.

SE를 이용하여 평균 앙상블의 정확도를 분석하였을 때 결과는 다음과 같다. 시나리오 수가 증가할수록 SE boxplot의 크기가 작아지나 그 평균은 1에 고정되어 있는 결과가 나타났다. 그리고 시나리오 간 상관성이 증가하면 boxplot의 크기가 커지고 시나리오 수 증가로 인한 boxplot의 크기 차이가 없어진다. 이는 평균 앙상블 시나리오를 이용하여 예측을 하였을 때 시나리오 수가 많을수록 더 안정적인 예측을 수행할 수 있으나, 상관성이 증가하면 정보의 손실로 인하여 많은 수의 시나리오를 사용하는 것의 효과가 없어진다는 것을 의미한다.

BS를 이용하여 ρ_a를 고려하지 않은 전체 결과를 분석하면 다음

과 같다. 시나리오 수가 증가할수록 BS의 평균이 감소하고 분산이 증가하는 결과가 나타났다. 그리고 시나리오 간 상관성이 증가하면 전체적으로 BS의 평균과 분산이 증가하며, 시나리오 수에 따른 BS 결과 차이가 감소한다. 이는 시나리오 수가 증가할수록 확률예측이 향상되지만 상관성이 증가할수록 정보 손실과 가능한 예측의 폭이 감소하여 극단적인 예측을 수행하게 된다는 것을 의미한다.

다음은 앙상블 예측의 정확도를 예상할 수 있는 상황인 ρ_a별 분석 결과이다. ρ_a별 SE 분석 결과는 다음과 같다. ρ_a가 0.1에서 0.9까지 증가하면서 전체적으로 오차가 줄어드는 결과가 나타났다. ρ_a가 0.1과 0.3인 경우에는 시나리오 간 상관성이 작을수록 그리고 시나리오 수가 많을수록 오차가 작아졌다. ρ_a가 0.5인 경우에서는 시나리오 간 상관성과 시나리오 수가 변하여도 SE 결과가 변하지 않는 현상이 나타난다. ρ_a가 0.7과 0.9인 경우에는 시나리오 간 상관성이 클수록 그리고 시나리오 수가 적을수록 오차가 작아지는 것을 확인할 수 있다.

ρ_a별 BS 분석 결과는 다음과 같다. ρ_a가 0.1, 0.3, 그리고 0.5인 경우에는 시나리오 간 상관성(ρ_e)이 작을수록 BS가 작게 나타난다. 그리고 시나리오 수가 많을수록 예측이 향상된다. ρ_a가 0.7인 경우에는 ρ_e의 크기와 예측의 정확도에 관계성이 나타나지 않는다. ρ_a가 0.9인 경우에는 앞의 경우와는 반대로 ρ_e가 클수록 BS가 작게 나타난다. ρ_e가 0.7과 0.9인 경우에만 시나리오 수의 증가가 예측의 정확도 향상에 도움이 되었으며, 나머지 ρ_e의 경우에서는 시나리오 수가 증가하면 BS가 커지는 결과가 나타났다.

시나리오 간 상관성이 앙상블 예측에 미치는 영향에 대한 결론은 다음과 같다. 앙상블의 예측의 정확도를 예상할 수 없는 경우에 시나리오 간 상관성이 작을수록 시나리오 수가 많을수록 더 좋은 예측을 수행할 가능성이 높아진다. 그리고 시나리오 간 상관성이 높을수록 정보의 손실로 인하여 오차가 증가하며, 다수의 시나리오

를 사용하는 효과가 감소한다.

ρ_a별로 분석 결과의 결론은 다음과 같다. 정보의 부족이나 입력 자료의 큰 불확실성으로 인하여 비교적 정확도가 낮은 예측이 예상될 경우에는 시나리오 수가 많고 상관성이 작을수록 더 정확도 높은 예측을 수행한다. 반대로 비교적 높은 정확도의 예측이 예상될 경우에는 상관성이 클수록 더 좋은 결과를 얻을 수 있다. 또한 시나리오 수의 증가가 오히려 예측의 정확도를 낮추는 결과를 보이기도 한다. 즉 높은 정확도의 예측이 보장되는 경우에는 앙상블 예측이 오히려 하나의 예측 시나리오만을 제공하는 예측 방법보다 더 낮은 예측 수행력을 보일 수 있다. 세 번째 상황은 양쪽 모두 아닌 경우이다. 높은 예측을 수행할 수 있거나 불확실성이 높아 낮은 예측밖에 수행할 수 없는 경우 외에 그 중간 정도의 정확도를 가지는 예측을 수행하는 상황에서는 시나리오 간 상관성이 앙상블 예측의 정확도에 영향을 미치지 못하는 결과가 나타난다. 그리고 확률예측으로써 앙상블을 사용한 경우에는 시나리오 수 증가가 상관성에 무관하게 예측의 정확도를 향상시킨다.

앙상블 시나리오 수가 앙상블 예측 시스템과의 관계를 알아보고 유효 앙상블 시나리오 수를 제시하기 위하여 2가지 방법을 사용하였다. 유효 앙상블 시나리오 수는 적절한 예측의 정확도를 가지는 최소한의 시나리오 수를 의미한다. 첫 번째 방법은 BS의 최대값과 최소값의 차이의 90% 되는 지점을 추정하여 개수를 정하였다. 그리고 두 번째 방법은 BS가 향상되는 정도를 그래프의 기울기로 판단하는 방법이다. 대부분의 경우에서 3개에서 5개로 시나리오 수가 늘어났을 때 가장 BS가 크게 향상되었다. 이때의 기울기의 5%가 되는 r구간을 유효 앙상블 시나리오 수 판단의 근거로 사용하였다. 전체 향상율의 90%가 되는 지점을 추정하였을 때 약 25개로 나타났고, 최대 기울기의 5%가 되는 구간을 추정하였을 때 대부분 15 ~ 20 구간에서 유효 시나리오 수를 추정할 수 있었다.

5.2 향후 연구

본 연구는 시나리오 간 상관성이 앙상블 예측에 미치는 영향을 분석하기 위하여 이론적인 실험을 수행하였다. 본 연구의 한계점과 향후 연구에 대하여 서술하면 다음과 같다.

앙상블 유량예측을 수행할 경우 크게 두 가지로 나누어 생각할 수 있다. 그것은 입력자료인 예측 강우와 강우-유출 모형이다. 우리나라에서 예측 강우는 MAPLE, KLAPS, 그리고 UM으로 세 종류가 있다. 시스템마다 오차의 구조가 다르며, 예측 강우의 분포가 서로 다를 수 있다. 마찬가지로 다양한 강우-유출 모형마다 오차의 구조와 예측 유량의 분포가 다를 수 있다. 본 연구에서는 상관성이 미치는 영향을 일반화하기 위하여 다변량 표준정규분포를 사용하였다. 이는 예측에 적용을 위하여 평균 예측값의 올바른 추정과 초기 조건에 대한 예측의 민감도가 중요하며, 정규분포가 이에 유용하다고 주장한 Stephenson and Doblas-Reyes(2000)의 연구에 근거하였다. 하지만 다른 분포를 따르는 다변량 상관 난수를 생성할 수 있다면 다양한 경우에 대하여 모의할 수 있을 것이다.

시나리오 간 상관성(ρ_e)은 0, 0.1, 0.3, 0.5, 0.7, 그리고 0.9으로 6 종류를 사용하였다. 이는 모두 양의 상관성으로 구성되어있다. 콜레스키 분해법을 사용하기 위해서는 행렬이 양의 양정치 행렬(positive-definite matrix)이 되어야 한다. 첨두시간에 대한 예측은 유량을 감소를 예측을 시나리오와 증가를 예측하는 시나리오 간의 작용으로 인하여 음의 상관계수가 나타날 수 있다. 따라서 첨두시간의 상관성을 보기 위해서는 앙상블 시나리오의 길이를 줄이거나 같은 증가율을 보이는 시나리오 간의 상관성을 분석할 필요가 있다.

실제로 유량 앙상블의 상관계수를 구해보면 하나의 상관행렬 내에 다양한 값이 나타난다. 가까운 초기시간을 가지는 시나리오 간에는 높은 상관성을 보이며, 예측시점이 비교적 떨어져있는 시나리

오 간에는 상관성이 크게 나타난다. 본 연구에서는 ρ_a별로 상관성이 앙상블 예측의 정확도에 미치는 영향이 다르게 나타났다. 앙상블 예측이 유용하게 사용될 수 있는 경우에 시나리오 간 상관성이 작을수록 더 높은 정확도의 예측을 수행할 가능성이 높아진다. 따라서 이 경우에 시나리오 간 상관성을 줄일 수 있는 방법이 필요하다. 확률예측의 경우 상관성을 고려하여 시나리오에 가중치를 적용하는 방법으로 시나리오 간 상관성을 감소시킬 수 있을 것으로 예상된다.

참고문헌

강부식, 유승엽, 고익환 (2007). “강수전망 가중앙상블을 이용한 장기 확률유량 예측.” **대한토목학회논문집**, 대한토목학회, 제27권, 제2-B호, pp. 183-191.

강태호 (2008). 앙상블 유량예측 시스템의 사전 및 사후처리에 관한 연구. 석사학위논문, 서울대학교.

강태호 (2010). "확률홍수예측의 필요성" **한국수자원학회논문집**, 한국수자원학회, 제43권, 제3호, pp. 75-82.

건설교통부, 한국수자원공사 (2004). 유역통합물관리를 위한 하천유출량 예측방안 연구.

건설교통부, 한국수자원공사 (2004). 한강유역조사 기본현황조사 보고서.

권현한, 문영일 (2004). “단일사상 강우-유출 모형의 불확실성 분석을 통한 홍수빈도곡선 유도(Ⅰ).” **대한토목학회논문집**, 대한토목학회, 제24권, 제3-B호, pp. 229-239.

기상청 (2009), 기상연감.

기상청, 국립기상연구소 예보연구과 (2009). 초단기 기상 분석 및 예측시스템(KLAPS) 개발 및 운영.

기상청, 국립기상연구소 예보연구과 (2009). MAPLE 적용 현황 및 평가.

김경진 (2011). 확률 홍수예측의 국내 도입을 위한 기초연구. 석사학위논문, 서울대학교.

김남원, 이정우, 김병식 (2009). 앙상블 모형을 이용한 확률적 유량예측 2차 보고서.

김동준, 박훈, 주상원, 임정옥, 박병권, 유희동, 조주영 (2009) “기상청 통합수치예보시스템의 예비성능평가.” **한국기상학회 봄 학술대회 논문집**, 한국기상학회, pp. 218-219.

김병식, 김보경, 권현한 (2009). “Meta-Gaussian 방법을 이용한 강우-유출 모형에서의 불확실성 산정.” **한국습지학회지**, 한국습지학회, 제11권, 제1호, pp. 49-64.

김보경, 장대원, 양동민, 유철상 (2009). “초단기 예측 모형 MAPLE 자료의 정확성 검토” **한국수자원학회논문집**, 한국수자원학회, 제42권, 제12호, pp. 52-64.

정대일, 김영오 (2002). “앙상블 예측을 이용한 충주댐의 월 유입량 예측.” **대한토목학회논문집**, 대한토목학회, 제22권, 제3B호, pp. 321-331.

오규창 (1998). 매개변수 추청을 통한 강우-유출 모형의 불확실성 분석. 박사학위논문, 서울대학교.

유원승, 김창환, 차준호 (2012). 수자원 현황과 전망.

이경태 (2012). 정선유역의 확률론적 홍수예측을 위한 강우 불확실성 분석. 석사학위 논문, 서울대학교.

이재경 (2005). 중장기 유량예측 향상을 위한 국내 기후정보의 이용. 석사학위논문, 서울대학교.

이재경, 김영오, 정대일 (2006). “중장기 유량예측 향상을 위한 국내 기후정보의 이용.” **한국수자원학회논문집**, 한국수자원학회, 제39-9호, pp. 755-766.

정대일, 김영오 (2002). “앙상블 예측을 이용한 충주댐 월 유입량 예측.” **대한토목학회논문집**, 대한토목학회, 제22권, 제3-B호, pp. 321-331.

정대일, 김영오, 조성준, 신현정 (2003). “앙상블 유량예측의 정확도 향상을 위한 강우-유출모형에 대한 연구: 1. 신경망 강우-유출모형.” **대한토목학회논문집**, 대한토목학회, 제23권, 제6-B호, pp. 521-530.

정대일, 이재경, 김영오 (2005). “최적선형보정을 이용한 앙상블 유량예측 시스템의 개선.” **한국수자원학회논문집**, 한국수자원학회, 제38권, 6호, pp. 471-483.

한강홍수통제소 (2005). 주요지천홍수예보보고서.

한강홍수통제소 (2008). 앙상블 모형을 이용한 확률적 유량예측 보고서.

Anderson, J. L. and Stern, W. F. (1995). "Evaluating the potential predictive utility of ensemble forecasts." *Journal of Climate*, Vol. 9, pp. 260-269.

Atger, F. (2001). "Verification of intense precipitation forecasts from single models and ensemble predcition systems." *Nonlinear Processes in Geophysics*, Vol. 8, pp. 401-417.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability." *Monthly Weather Review*, Vol. 78, No. 1, pp. 1-3.

Brnkovic, C. and Palmer, T. N. (1997). "Atmoshperic seasonal predictability and estimates of ensemble size." *Monthly Weather Review*, Vol. 125, pp. 859-874.

Buizza, R. (1997). "Potential forecast skill of ensemble prediction and spread and skill distribution of the ECMWF ensemble prediction system." *Monthly Weather Review*, Vol. 125, pp. 99-119.

Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, J., Hamrud, M, and Hollingsworth, A. (1998). "Impact of model resolution and ensemble size on the performance of an ensemble prediction system." *Quarterly Journal of the Royal Meteorological Society*, Vol. 124, pp. 1935-1960.

Choi, H. T. and Beven, K. J. (2007). "Multi-period and multi-criteria model condtioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework." *Journal of Hydrology*, Vol. 332, pp. 316-336.

Cloke, H. L. and Pappenberger, F. (2009). "Ensemble flood forecasting: A review." *Journal of Hydrology*, Vol. 375, pp. 613-626.

Day, G. N. (1985). "Extended streamflow forecasting using NWSRFS." *Journal of Water Resources Planning and Management*, Vol. 111, No. 2, pp. 147-170.

Demeritt, D. et al. (2007). "Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting." *Environmental Hazard,* Vol. 7, No. 2, pp. 115

Ferro, C. A. T. (2007). "Comparing probabilistic forecasting systems with the Brier score." *Weather and Forecasting*, Vol. 22, pp. 1076-1088.

Fleishman, A. (1978). "A method for simulating non-normal distributions." *PSYCHOMETRIKA*, Vol. 42, No. 4, pp. 521-532.

Fread, D. L. (1998). "A perspective on hydrologic prediction trends." *Special Symposium on Hydrology*, AMS, Phoenix, Arizona, pp. j1-j6.

Graygier, J. G. and Stedinger, J. R. (1989). "A generalized maintenace of variance extension procedure for extending correlated series." *Water Resources Research*, Vol. 25, No. 3, pp. 345-349.

Hamill, T. M. and Colucci, S. J. (1996). "Verification of Eta-RSM short-range ensemble forecasts." *Monthly Weather Review*, Vol. 125, pp. 1312-1327.

Heckman, J. J., Schmierer, D., and Urzua, S. (2010). "Testing the correlated random coefficient model." *Journal of Econnometrics*, Vol. 158, pp. 177-203.

Hoffman, P. J. (1959). "Generating variables with arbitrary properties."

PSYCHOMETRIKA, Vol. 24, No. 3, pp. 265-267.

Ingram, J. J., Hudlow, M. D. and Fread, D. L. (1995). "Hydrometeorological coupling for extended streamflow predictions." *Proceedings of Conference on Hydrology, 75th Annual Meeting of AMS*, American Meteorological Society, Boston, Mass., pp. 186-191.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schar, C. (2008). "A probabilistic view on the August 2005 floods in the upper Rhine catchment." *Natural Hazards Earth System Sciences*, Vol. 8, pp. 281-291.

Jones, M. S. and Colle, B. A. (2006). "Evaluation of mesoscale short-range ensemble forecast system over the northest United States." *Weather and Forecasting*, Vol. 22, pp. 36-55.

Kaiser, H. F. and Dickman, K. (1962). "Sample and poluation score matrices and sample correlation matrices from anarbitrary population correlation matrix." *PSYCHOMETRIKA*, Vol. 27, No. 2, pp. 179-182.

Kim, Y.-O., and Jeong D. I. (2004). "Ensemble streamflow prediction using climate forecast information." *ICHWC 2004*, Seoul, Korea, CD.

Kim, Y.-O., Jeong, D. I., Kim, H. S. (2001). "Improving water supply outlooks in Korea with ensemble streamflow prediction." *Water International*, Vol. 26, No. 4, pp. 563-568.

Lunn, A. D. and Davies, S. J. (1998). "A note on generating correlated binary variables." *Biometrika*, Vol. 85, No. 2, pp. 487-490.

Lane, W. L. and D. K. Frevert (1988). "Applied stochastic techniques:

LAST computer package: user manual."

Lurie, P. M. and Goldberg, M. S. (1998). "An approximate method for sampling correlated random variables from partially-specified distributions." *Management Science*, Vol. 44, No. 2, pp. 203-218.

Lurie, P. M. and Goldberg, M. S. (1998). "An approximate method for sampling correlated random variables from partially-specified distributions." *Management Science*, Vol. 44, No. 2, pp. 203-218.

Madsen, H. and D. Rosbjerg (1997). "Generalized least squares and empirical bayes estimation in regional partial duration series index flood modeling." *Water Resources Research,* Vol. 33, No. 4, pp. 771-781.

Madsen, H., Mikkelsen, P. S., Rosbjerg, D., Harremoës, P. (2002). "Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics." *Water Resources Research*, Vol. 38, No. 11.

Maidment, D. R. (1992). Handbook of Hydrology, McGraw-Hill Inc.

Mejia, J. M. and J. Rousselle (1976). "Disaggregation models in hydrology revisited." *Water Resources Research,* Vol. 12, No. 2, pp. 185-186.

Montanari, A. (2005). "Large sample behaviors of the generalized likelihood uncertainty estimation(GLUE) in assessing the uncertainty of rainfall-runoff simulations." *Water Resources Research*, Vol. 41, pp. 1-13.

Murphy, A. H. (1988). "Skill score and correlation coefficients in model verification." *Monthly Weather Review*, Vol. 117, pp.

572-581.

Murphy, J. M. (1988). "The impact of ensemble forecasts on predictability." *Quarterly Journal of the Royal Meteorological Society*, Vol. 114, pp. 463-493.

Murphy, J. M. (1990). "Assessment of the practical utility of extended range ensemble forecasts." *Quarterly Journal of the Royal Meteorological Society*, Vol. 116, pp. 89-125.

Najafi, M. R., Moradkhani, H., and Piechota, T. C. (2012). "Ensemble streamflow prediction: climate signal weighting methods vs. climate forecst system reanalysis." *Journal of Hydrology*, Vol. 442-443, pp. 105-116.

Ouarda, T. B., Girard, C., Cavadias, G. S., and Bobée, B. (2001). "Regional flood frequency estimation with canonical correlation analysis." *Journal of Hydrology,* Vol. 254, No. 1, pp. 157-173.

Pappenberger, F. and Beven, K. (2004). "Functional classification and evaluation of hydrographs based on multicomponent mapping." *International Journal of River Basin Management*, Vol. 2, No. 1, pp. 1-8.

Parker, D. and M. Fordham (1996). "An evaluation of flood forecasting, warning and response systems in the European Union." *Water Resources Management,* Vol. 10, No. 4, pp. 279-302.

Pekárová, P., Miklánek, P., and Pekár, J. (2003). "Spatial and temporal runoff oscillation analysis of the main rivers of the world during the 19th - 20th centuries." *Journal of Hydrology,* Vol. 274, No. 1, pp. 62-79.

Pitt, M. and G. Britain (2008). Learning lessons from the 2007 floods, Cabinet Office, London

Rencher, A. C. and W. F. Christensen (2012). Methods of multivariate analysis ,John Wiley & Sons.

Roulston, M. S. (2007). "Performance targets and the Brier score." *Meteological Applications*, Vol. 14, pp. 185-194.

Saad, M. and A. Turgeon (1988). "Application of principal component analysis to long term reservoir management." *Water Resources Research*, Vol. 24, No. 7, pp. 907-912.

Schakke, J. C. (1973). "Disaggregation processes in stochastic hydrology." *Water Resources Research,* Vol. 9, No. 3 pp. 580-585.

Schaake, J. C., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., and Seo, D. J. (2007). "Prediction and temperature ensemble forecasts from single-value forecasts." *Hydrologic Earth System Sciences Discuss,* Vol. 4, pp. 655-717.

Smith, J. A., Day, G. N., and Kane, M. D. (1992). "Nonparametric framework for long-large streamflow forecasting." *Journal of Water Resources Planning and Management*, ASCE, Vol. 118, No. 1, pp. 82-92.

Stedinger, J. R., and Tasker, G. D. (1985). "A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations." *Water Resources Research*, Vol. 21, No. 5 pp. 665-675.

Stedinger, J. R. and Y.-O. Kim (2010). "Probabilities for ensemble forecasts reflecting climate information." *Journal of Hydrology,* Vol. 391, No. 1 pp. 9-23.

Stephenson, D. B. and Doblas-Reyes, F. J. (2000). "Statistical methods for interpreting Monte Carlo ensemble forecasts." *Tellus*, Vol. 52A, pp. 300-322.

Stewart, I. T., Caya, D. R., Dettinger, M. D. (2005). "Changes toward earlier streamflow timing across western North America." *Journal of Climate,* Vol. 18, No. 8 pp. 1136-1155.

Tadikamalla, P. R. (1980). "On simulating non-normal distributions." *PSYCHOMETRIKA*, Vol. 45, No. 2, pp. 273-279.

Tsakiris, G., Nalbantis, I., and Cavadias, G. (2011). "Regionalization of low flows based on canonical correlation analysis." *Advances in Water Resources*, Vol. 34, No. 7, pp. 865-872.

Touran, A. and Wiser, E. P. (1992). " Monte Carlo techinque with correlated random variabels." *Journal of Construction Engineering and Management,* Vol. 118, pp. 258-272.

Wherry, R. J., Naylor, J. C., and Fallis, R. F. (1965). "Generating multiple samples of multivariate data with arbitrary population parameters." *PSYCHOMETRIKA*, Vol. 30, No. 3, pp. 303-313.

Western, A. W., Grayson, R., McMahon, T., Blöschl, G., and Wilson, D. J. (2004). "Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes." *Journal of Hydrology,* Vol. 286, No. 1, pp. 113-134.

Xu, C.-y., Gong, L., Jiang, T., Chen, D., and Singh, V. (2006). "Analysis of spatial distribution and temporal trend of reference evapotranspiration and pan evaporation in Changjiang (Yangtze River) catchment." *Journal of Hydrology,* Vol. 327, No. 1, pp. 81-93.

Abstract

Effects of Correlated Ensemble Scenarios on Accuracy of Probabilistic Flood Forecasting

Young-Ho Seo
Department of Civil & Environmental Engineering
The Graduate School
Seoul National University

Flood forecasting systems around the world have continued to utilize Ensemble Streamflow Prediction (ESP) over the single deterministic forecasts for operational and research fields (Cloke and Pappenberger, 2009). The uncertainty of probabilistic forecasting should be studied to represent the probabilisitc occurrence of events.

It is assumed that Many ESP are comprised of equiprobable ensemble members. Therefore, ESP should have an adequate number of ensemble members to describe the full range probabilities of streamflow (Cloke and Pappenberger, 2009). It is very difficult to determine the reasonable number of scenarios in ensemble prediction. There are several ESP research related to the number of ensemble members. However, it is still difficult to suggest the criteria to establish the adequate number of scenarios.

The assumption of equiprobability implies that ensemble scenarios are uncorrelated. The existence of correlation between ensemble members can cause increasing uncertainty and decreasing reliability, since the correlation between the variables generally indicates loss of information.

In this study, the effective number of the ensemble members is defined

as a minimum number that can maintain an acceptable accuracy level of ESP. Therefore, this study handles two issues related to ESP; one is the cross-correlation between ensemble members and the other is the number of ensemble members.

It is assumed that each scenario follows the Gaussian distribution in order to draw general conclusions. This study tested 6 cross-correlation values(ρ_e = 0, 0.1, 0.3, 0.5, 0.7, 0.9) and 10 ensemble members(p = 3, 5, 7, 9, 12, 15, 20, 30, 50, 100). To analyze the effects of correlation on the ensemble, the generation of observed scenario are necessary to evaluate the simulation results. Therefore, by generating the observed scenario with different values of 'nominal' accuracy, the effects of correlation and the number of ensembles can be determined.

Brier score is used to evaluate the probabilistic forecasts. D^2 and SE are evaluation index for deterministic forecasts. In case of inaccurate forecasts, the accuracy of ESP is improved as ensemble correlation decreases or the number of ensemble mebers increases. In contrast, when very accurate forecasts are available, there are limited benefits from utilizing ESP.

Two methods were developed to determine the effective number of the ensemble members. In the first method, the difference between the largest and smallest Brier score was determined in a Brier score curve and used to define the effective number. The second method involved comparison of the slope between each Brier score intervals. As a results, 20~25 scenarios are suitable for the effective number regardless of the ensemble correlation

Keywords: ensemble streamflow prediction, correlation, multivariate correlated random number generation, the effective number, Brier score

Student Number: 2012-20904