Post-processing multimodel seasonal climate forecasts: When can we expect an improvement from unequal weighting, and how big is the expected improvement?

Stefan Siegert
September 1, 2016

1 Introduction

We usually have multiple ensembles for the same prediction target, i.e., an ensemble of ensembles, sometimes also called a superensemble. Multimodel ensembles (MMEs) are highly structured data with a rich correlation structure, i.e., different models have different correlations with the observations, and models also are positively correlated with one another. An important questions is how the various models should be optimally combined into a single prediction for future observations.

It has been pointed out by previous studies that weighting all the models equally is often preferable, even when we have reason to believe that the weights should be different. The reason is that there is only a finite amount of data available to estimate the combination weights. Random estimation errors of the combination weights degrade the quality of the post-processed forecast, leading to post-processed forecasts that perform worse than weighting the forecasts equally.

- DelSole (2013): Is unequal weighting significantly better than equal weighting for multi-model forecasting? (QJ 139, 176–183)
- Weigel et al (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? QJ 134:241–260 (2008)
- Weigel et al (2010) Risks of model weighting in multimodel climate projections. JCLIM 4175ff

In this note, we address questions in this context that have not been addressed in previous studies, namely:

- How should we optimally combine the MME into a single prediction for the observation? In particular, should all forecasts be weighted equally, or

should we assign different weights to different forecasts?
- When can we expect unequal weighting to be beneficial, and how big is the expected improvement?

2 Seasonal forecast data: El Nino Southern Oscillation and the North Atlantic Oscillation

For illustration, we use seasonal forecasts of El-Niño Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO). The ENSO hindcasts are for the time period 1982–2010 ($n = 29$) using model runs that are initialised in August each year, forecasting the Nino3.4 index for December (5 months lead time). The multi-model ensemble (MME) uses forecast runs from ECMWF System4 ($M = 51$ members), NCEP CFSv2 ($M = 24$), MeteoFrance System3 ($M = 11$), GFDL ($M = 10$), CMC2 ($M = 10$), and NASA ($M = 10$). The NAO hindcasts are for the time period 1975–2002 ($n = 28$) using runs that are initialised in November each year, forecasting average NAO for the following December-February (1-3 months lead time). The MME uses data from ECMWF, LODYN, MeteoFrance, MPI, and UK MetOffice that was generated in the DEMETER project. Forecast and observation data are shown in Figure 1.

Below we summarise means, variances (using $1/n$ instead of $1/(n-1)$), and correlations of ENSO and NAO data:

ENSO statistics	obs	cfs	cmc	gfdl	mf	nasa	ec
mean	26.70	25.64	25.77	24.89	27.05	25.29	24.82
stdev	1.21	1.41	1.54	1.30	0.67	1.19	1.24

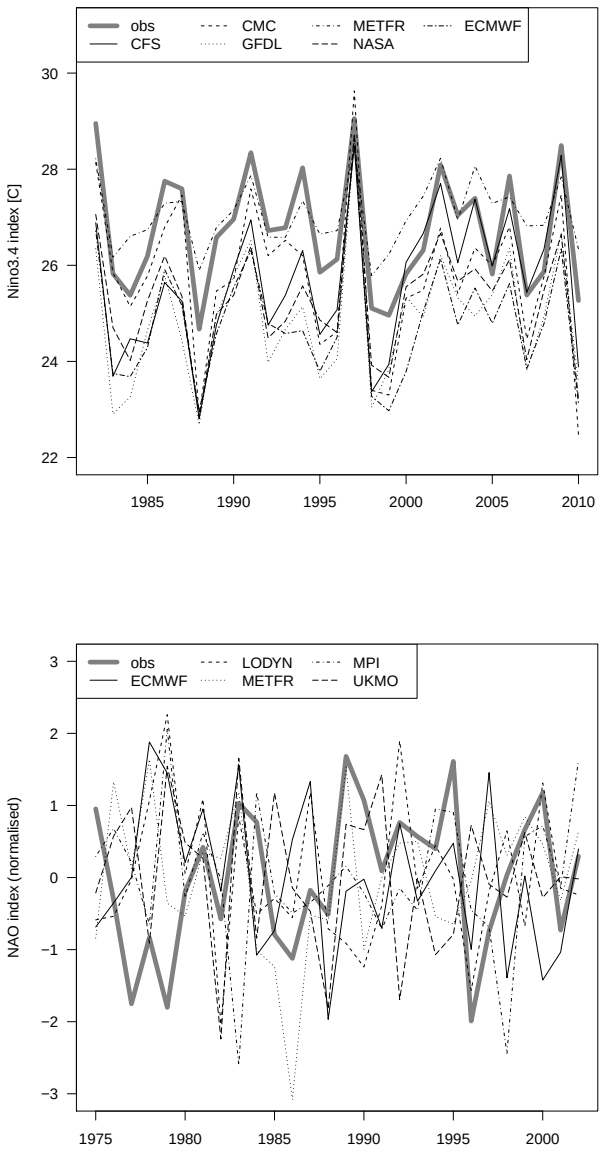


Figure 1: ENSO and NAO multi-model forecasts and observations.

ENSO correlations	obs	cfs	cmc	gfdl	mf	nasa	ec
obs	1.00	0.81	0.89	0.85	0.87	0.88	0.91
cfs		1.00	0.78	0.89	0.93	0.89	0.86
cmc			1.00	0.82	0.83	0.90	0.91
gfdl				1.00	0.87	0.93	0.92
mf					1.00	0.90	0.90
nasa						1.00	0.94
ec							1.00

NAO statistics	obs	ecmwf	lodyn	metfr	mpi	ukmo
mean	0.6	0.75	0.70	1.03	0.74	0.60
stdev	1.4	0.14	0.12	0.16	0.10	0.14

NAO correlations	obs	ecmwf	lodyn	metfr	mpi	ukmo
obs	1.00	-0.070	0.03	0.18	0.019	-0.14
ecmwf		1.000	0.47	0.23	-0.006	0.14
lodyn			1.00	0.16	0.051	0.06
metfr				1.00	-0.130	0.09
mpi					1.000	-0.10
ukmo						1.00

3 Optimal combination by multivariate analysis

For statistical post-processing, it is necessary to make distributional assumptions about the joint behavior of forecasts and observations. We will work under the assumption that forecasts and observations are jointly Normally distributed. In particular, we assume that the observation y_t and the vector of forecasts $\boldsymbol{x}_t =$

$(x_t^{(1)}, \ldots, x_t^{(M)})'$ at time t are jointly drawn from a multivariate Normal distribution with fixed parameters:

$$\begin{bmatrix} y_t \\ \boldsymbol{x}_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right). \tag{1}$$

Draws at different times are assumed to be independent. Any biases in mean and scale, and varying levels of correlation between forecasts and observations are encoded in the paramters μ and Σ of the joint Normal distribution. These parametric assumptions allow us to derive the optimal post-processing of ensemble means into a prediction for the observation.

Under the assumption of joint Normality, the conditional distribution of the observation, given specific values of the forecasts $\boldsymbol{x} = \boldsymbol{m}_t = (m_t^{(1)}, \ldots, m_t^{(M)})'$, is a Normal distribution

$$p(y_t | \boldsymbol{x}_t = \boldsymbol{m}_t) = \mathcal{N}(\mu_{y|x}(\boldsymbol{m}_t), \Sigma_{y|x}) \tag{2}$$

with conditional mean $\mu_{y|x}$ and conditional variance $\Sigma_{y|x}$ given by

$$\mu_{y|x}(\boldsymbol{m}_t) = \mu_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{m}_t - \boldsymbol{\mu}_x) \tag{3}$$

$$\Sigma_{y|x} = \Sigma_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \tag{4}$$

(c.f. Mardia et al 1995 ch 3). From eq. 3 it follows that the optimal combination of the M forecasts $(m_t^{(1)}, \ldots, m_t^{(M)})$ into a single prediction $\hat{y}_t$ for the observation, is given by the linear function

$$\hat{y}_t = a + \sum_{m=1}^{M} w_m x_t^{(m)} \tag{5}$$

where

$$a = \mu_y - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\mu}_x \tag{6}$$

$$w_m = \boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx}^{-1})_m. \tag{7}$$

Here $(\boldsymbol{\Sigma}_{xx}^{-1})_m$ denotes the mth column of the inverse of $\boldsymbol{\Sigma}_{xx}$. Eq. 7 shows that the weight of forecast m in the combined prediction depends on the covariances of all forecasts with the observations, as well as the covariances of forecast m with all the other forecasts.

In general the parameters of the Normal distribution are unknown, and have to be estimated from available tuples of past forecasts and observations. We will look into the problem of parameter estimation later. For the moment, we assume that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known exactly. Plugging in these values into equations 6 and 7 then yields the overall mean a, the post-processing weights w_m, and the residual variance $\Sigma_{y|x}$. If we assume that the mean and covariance of the joint distribution of forecasts and observations are equal to the sample means and covariances of the ENSO data, we get

a	w_{cfs}	w_{cmc}	w_{gfdl}	w_{mf}	w_{nasa}	w_{ec}	$\Sigma_{y\|x}$
-4.07	-0.10	0.33	0.17	0.66	-0.16	0.28	0.19

and if we plugin the sample means and covariances of the NAO data, we get

a	w_{ecmwf}	$\mathrm{w}_{\mathrm{lodyn}}$	$\mathrm{w}_{\mathrm{metfr}}$	$\mathrm{w}_{\mathrm{mpi}}$	$\mathrm{w}_{\mathrm{ukmo}}$	$\Sigma_{y\|x}$
-0.37	-1.27	0.74	1.90	0.38	-1.41	1.58

For simplicity, it is often preferred to combine models from different forecast centres by a democratic vote, where each model gets the same weight in the combination. Such an equal-weights post-processingn scheme is equivalent to post-processing the multi-model mean (MMM), by first combining all forecasts into the MMM, and then fitting the observation to the MMM by linear regression.

Given the above correlation structures, what would be the optimal post-processing scheme for the multi-model mean (MMM), i.e., for post-processing with equal weights? We use the well-known result from Normal theory: If the vector $\boldsymbol{x}$ is a random variable $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the linear transformation $\boldsymbol{A}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$. We still assume the multivariate Normal structure of eq. 1. To combine the individual forecasts into the MMM, we set

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \frac{1}{M} & \dots & \frac{1}{M} \end{pmatrix} \tag{8}$$

We can calculate the joint distribution of the observation and the MMM. From the joint distribution, we obtain the forecast distribution $p(y_t|\bar{x}_t = m_t)$ using eq. 2:

$$p(y_t|\bar{x}_t = \bar{m}_t) = \mathcal{N}\left(\mu_y + \frac{\Sigma_{y\bar{x}}}{\Sigma_{\bar{x}\bar{x}}}(\bar{m}_t - \mu_{\bar{x}}), \Sigma_{yy} - \frac{\Sigma_{\bar{x}y}^2}{\Sigma_{\bar{x}\bar{x}}}\right) \tag{9}$$

where $\Sigma_{y\bar{x}}$, $\Sigma_{\bar{x}\bar{x}}$ and $\mu_{\bar{x}}$ are given by

$$\mu_{\bar{x}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\mu}_x^{(m)} \tag{10}$$

$$\Sigma_{\bar{x}\bar{x}} = \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} \boldsymbol{\Sigma}_{xx}^{(m,m')} \tag{11}$$

$$\Sigma_{y\bar{x}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\Sigma}_{xy}^{(m)} \tag{12}$$

Post-processing with equal weights is a simple linear regression with parameters a, w_{MMM} and $\Sigma_{y|x}$ as in eq. 5. Assuming the ENSO and NAO correlation structures, we get the following equal-weighting post-processing parameters

	a	w_{MMM}	$\Sigma_{y\|x}$
ENSO	2.32	0.95	0.24
NAO	0.34	0.34	1.70

Lastly, we consider the trivial post-processing scheme where the forecast model values are ignored, and a climatological forecast is made for the observation. The climatological forecast is equivalent to a post-processing with zero weights, i.e.

$$\mu_{y|x} = \mu_y \tag{13}$$

$$\Sigma_{y|x} = \Sigma_{yy} \tag{14}$$

We consider the questions what is the potential benefit of using unequal weighting vs equal weighting for MMEs with a known correlation structure. To address this, we calculate expected verification scores, in particular the expected squared error

of the forecast mean (SQERR), the continuous ranked probability score (CRPS) and the logarithmic score (LOGS). The corresponding equations are

$$SQERR(\mathcal{N}(\mu_{y|x}, \Sigma_{y|x}), y) = (\mu_{y|x} - y)^2 \tag{15}$$

$$CRPS(\mathcal{N}(\mu_{y|x}, \Sigma_{y|x}), y) = \sigma_{y|x}(2\varphi(z) + z(2\Phi(z) - 1) - \pi^{-\frac{1}{2}}) \tag{16}$$

$$LOGS(\mathcal{N}(\mu_{y|x}, \Sigma_{y|x}), y) = \frac{1}{2}(\log 2\pi + \log \Sigma_{y|x} + z^2) \tag{17}$$

where $\sigma_{y|x} = \sqrt{\Sigma_{y|x}}$, $z = (y - \mu_{y|x})/\sigma_{y|x}$, and φ and Φ are the pdf and cdf of the standard Normal distribution. The expectations of the above scores can be calculated analytically under the joint Normal assumption.

In all three post-processing scenarios (climatology, equal weights, unequal weights), the observation y_t is assumed to have a Normal distribution $\mathcal{N}(\mu, \sigma^2)$, with constant variance, and a mean that is either constant (climatological forecast) or depends on the value of the forecasts (equal and unequal weighting). To calculate the expected value of the score $S(\mathcal{N}(\mu, \sigma^2), y)$, we use conditional expectations:

$$E_{x,y}S\left[\mathcal{N}(\mu_{y|x}, \Sigma_{y|x}), y\right] = E_x\left[E_{y|x}S\left[\mathcal{N}(\mu_{y|x}, \Sigma_{y|x}), y\right]\right] \tag{18}$$

The inner expectation, taken over all possible values of the observation y for a given value of the forecast x, is given by

$$E_{y|x}\mathrm{SQERR} = \Sigma_{y|x} \tag{19}$$

$$E_{y|x}\mathrm{CRPS} = \sqrt{\frac{\Sigma_{y|x}}{\pi}} \tag{20}$$

$$E_{y|x}\mathrm{LOGS} = \frac{1}{2}\left(\log 2\pi + \log \Sigma_{y|x} + 1\right) \tag{21}$$

None of these expected scores depend on the particular value of x, because the forecast variance is constant in all three forecasting schemes. Taking the expectation E_x therefore leaves the expressions in eq. 19-21 unchanged.

The expected score values assuming the ENSO and NAO correlation structures are:

ENSO	SQERR	CRPS	LOGS
clim	1.51	0.69	1.63
equal	0.24	0.28	0.71
unequal	0.19	0.25	0.59
NAO	SQERR	CRPS	LOGS
clim	1.9636	0.79058	1.7563
equal	1.9630	0.79046	1.7562
unequal	1.8244	0.76205	1.7196

For both ENSO and NAO, the unequal weighting performs better than equal weighting on average. This result is expected because the correlations of the different models with the observation are different. Since we are working under assumption that all distributional parameters are known exactly, equal weighting leads to a worse forecast than unequal weighting.

The analysis can be used to get a feeling of the magnitudes of improvement that we might hope to get from using the optimal weighting with unequal weights as opposed to using the sub-optimal weighting with equal weights. For ENSO, the improvement of equal weighting over the climatology is very large for all scores, compared to the improvement of unequal weighting over equal weighting. For NAO, all improvements in scores are tiny; but the improvement of unequal weighting over equal weighting is much larger than the improvement of equal weighting over the climatological forecast.

The improvements can be quantified by relative scores which we define here as the score differences between equal and unequal weighting, normalised by the score difference between climatological and unequal weighting:

$$Skill = \frac{S_{equal} - S_{unequal}}{S_{clim} - S_{unequal}} \tag{22}$$

The skill scores are given by

	SQERR	CRPS	LOGS
ENSO	0.038	0.069	0.113
NAO	0.9957	0.9958	0.9959

The relative improvements are vastly different between ENSO and NAO. For ENSO, the scores achieved by equal weighting is relatively close the scores achieved by unequal weighting, but for NAO the equal weighting scores are (relatively) rarely better than the climatological scores. The behavior of the different scores is similar: The log score indicates the biggest expected relative improvement of unequal weighting over equal weighting, and the squared error the smallest.

The above analysis was based on highly idealised assumptions. In general we do not know the parameters of the joint distribution of forecasts and observations exactly. In general all we have is a finite sample of forecasts and observations. We can make an assumption of joint Normality and independence and estimate the parameters. In this case, the quality of our post-processed forecasts will also be affected by random errors due to the finiteness of the data set that is used for estimating the parameters.

4 Equal vs unequal weighting based on finite training data

We set up a random number experiment to study the effect of finite hindcast size on how well the post-processing weights can be estimated. We fix parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a multivariate Normal distribution, from which we draw independent samples that represent forecasts and observations. We draw a training data set of size n, representing n time samples of forecasts and observations. From the training data set, we estimate the mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ which we use to calculate the weights to post-process future forecasts. Then we draw one more sample from the same multivariate distribution as a test case, i.e., we get one additional set of forecasts $\boldsymbol{x}_{n+1}$ and a single verifying observation y_{n+1}. We use the post-processing parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ to transform the forecasts $\boldsymbol{x}_{n+1}$ into a predictive distribution for the observation y_{n+1}, using eq **??** for unequal weighting, and eq. **??** for equal weighting. The forecast is evaluated by the squared error of the mean, the logarithmic score and the CRPS. For every setting of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and n, this procedure is repeated 10^6 times, resulting in a collection of 10^6 values of verification scores, that evaluate the of out-of-sample predictions.

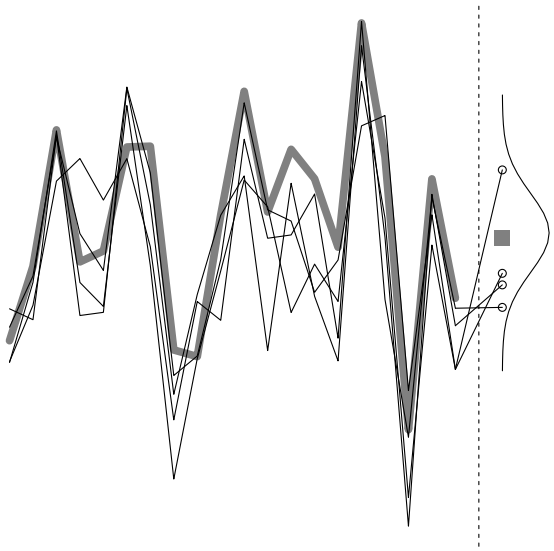


Figure 2: Illustration of the random number experiment. The black and gray solid lines to the left of the dashed lines are ensemble means and observations in the training data set. These data are used to estimate the post-processing weights. A single set of forecasts is produced out of sample (open circles to the right of the dashed line), and post-processed into a predictive distribution function, using the parameters estimated from the training data set. This forecast distribution is compared with the out-of-sample verifying observation (gray square) using the squared error, the CRPS, and the Log score.

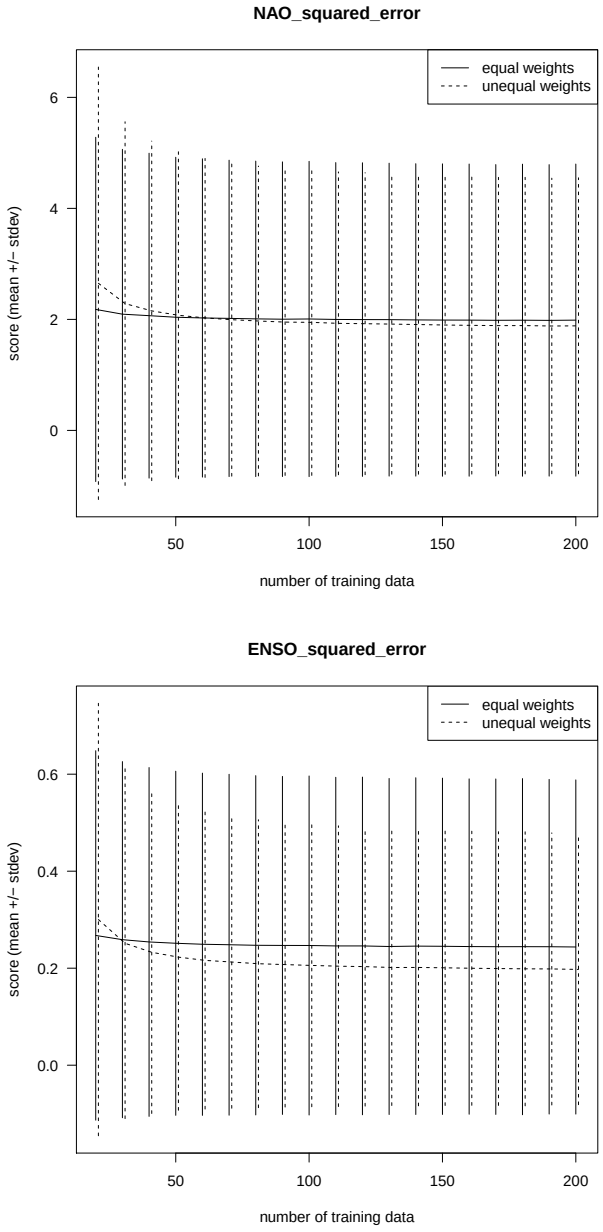


Figure 3: Test

Figure 3 shows how the mean and standard deviation of the squared prediction error behaves as a function of the training sample size n. To set the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we used the sample values calculated for the NAO and ENSO seasonal prediction data sets. The standard deviations are calculated over all out-of-sample predictions to indicate the range of score values that can be expected for any given forecast. We note the following:

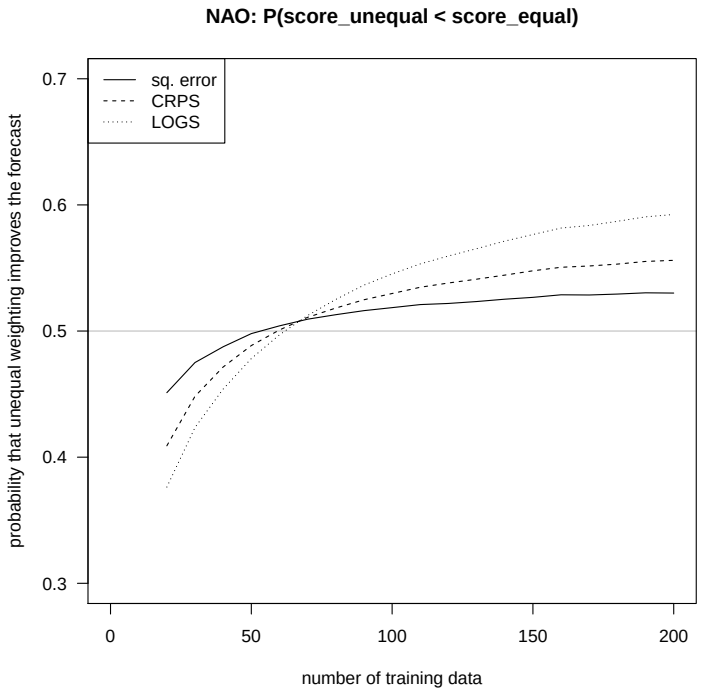
1. For small sample sizes, the expected squared error of the forecast mean is smaller if we use equal weights than if we use unequal weights. For the ENSO setting, equal weighting is better if the training sample size $n \leq 30$, and for the NAO setting, equal weighting is better on average if $n \leq 60$.
2. For large training sample sizes, unequal weighting outperforms equal weighting on average. But the magnitude of the improvement is very small compared to the total variability of verification scores.
3. The above points are true for the logarithmic score and CRPS as well.

The above observations raise a number of questions:

1. Given the small number of hindcast samples that we normally have for seasonal predictions, should we always use equal weighting instead of unequal weighting?
2. Given the small effect that we can expect from unequal weighting, is it worthwhile to pursue this form of post-processing at all?
3. Given the large variability of values of the verification score, how likely are we to improve the forecasts by unequal weighting on any particular forecast instance?

The first two questions are subjective and have to judged on a case by case basis, perhaps using simulation studies as we did above. The third question can be answered for our simulation study. For each foreacst instance, we get two post-processed forecasts for the same observation, one based on equal weighting and one based on unequal weighting. Each forecast obtains a certain value of the verification score. On any particular forecast instance, equal weighting is either better or worse than unequal weighting. We can now ask the questions "On what fraction of forecast instances does unequal weighting produce a better forecast than equal weighting"?

In Figure 4, the fraction of cases where unequal weighting produces a better fore-


NAO: P(score_unequal < score_equal)
sq. error
CRPS
LOGS
probability that unequal weighting improves the forecast
0.3
0.4
0.5
0.6
0.7
0
50
100
150
200
number of training data

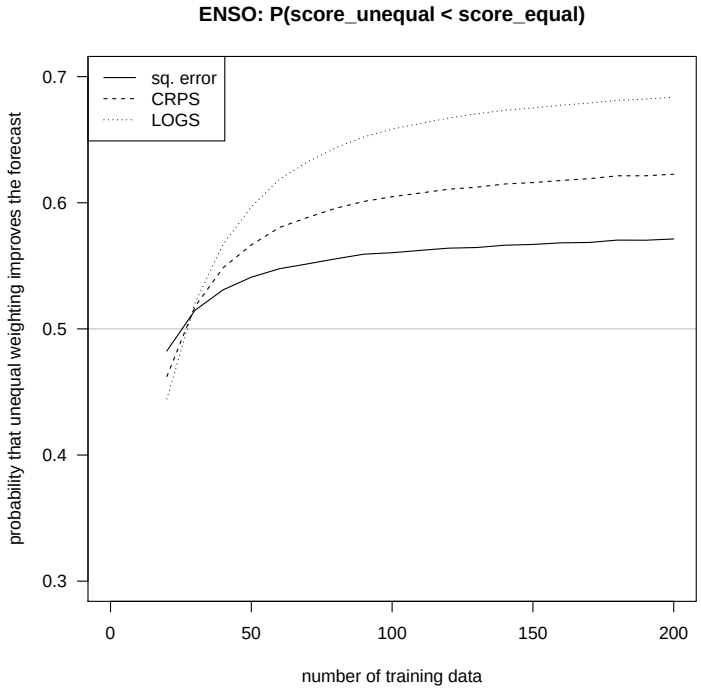

ENSO: P(score_unequal < score_equal)
sq. error
CRPS
LOGS
probability that unequal weighting improves the forecast
0.3
0.4
0.5
0.6
0.7
0
50
100
150
200
number of training data

Figure 4:

cast than equal weighting is shown as a function of the training sample size n. We find that for small sample sizes, equal weighting is more likely than not to produce a better forecast than unequal weighting. For larger sample sizes, unequal weighting becomes more likely to perform better. But even for large training sample sizes, there is still a considerable number of cases, where equal weighting outperforms unequal weighting. Furthermore, the 3 scoring rules considered behave rather differently. The logarithmic score is much more likely to show an improved forecast due to unequal weightingthan the other two scores. The mean squared error is least likely to indicate an improvement of unequal over equal weighting.