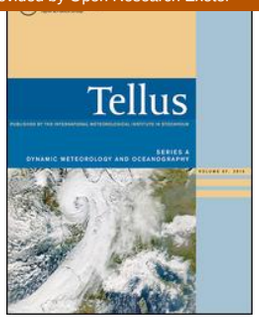

Tellus A: Dynamic Meteorology and Oceanography

ISSN: (Print) 1600-0870 (Online) Journal homepage: https://www.tandfonline.com/loi/zela20

Predictive verification for the design of partially exchangeable multi-model ensembles

Zied Ben Bouallégue, Christopher A. T. Ferro, Martin Leutbecher & David S. Richardson

To cite this article: Zied Ben Bouallégue, Christopher A. T. Ferro, Martin Leutbecher & David S. Richardson (2020) Predictive verification for the design of partially exchangeable multi-model ensembles, Tellus A: Dynamic Meteorology and Oceanography, 72:1, 1-12, DOI: 10.1080/16000870.2019.1697165

To link to this article: https://doi.org/10.1080/16000870.2019.1697165

Published online: 18 Dec 2019.

Tellus
SERIES A
DYANAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Predictive verification for the design of partially exchangeable multi-model ensembles

By ZIED BEN BOUALLÉGUE[1]*, CHRISTOPHER A. T. FERRO[2], MARTIN LEUTBECHER[1], and DAVID S. RICHARDSON[1], [a]*The European Centre for Medium-Range Weather Forecasts, ECMWF, United Kingdom;* [b]*Department of Mathematics, University of Exeter, United Kingdom*

(Manuscript received 1 February 2019; in final form 30 September 2019)

ABSTRACT

The performance of an ensemble forecast, as measured by scoring rules, depends on its number of members. Under the assumption of ensemble member exchangeability, ensemble-adjusted scores provide unbiased estimates of the ensemble-size effect. In this study, the concept of ensemble-adjusted scores is revisited and exploited in the general context of multi-model ensemble forecasting. In particular, an ensemble-size adjustment is proposed for the continuous ranked probability score in a multi-model ensemble setting. The method requires that the ensemble forecasts satisfy generalised multi-model exchangeability conditions. These conditions do not require the models themselves to be exchangeable. The adjusted scores are tested here on a dual-resolution ensemble, an ensemble which combines members drawn from the same numerical model but run at two different grid resolutions. It is shown that performance of different ensemble combinations can be robustly estimated based on a small subset of members from each model. At no additional cost, the ensemble-size effect is investigated not only considering the pooling of potential extra-members but also including the impact of optimal weighting strategies. With simple and efficient tools, the proposed methodology paves the way for predictive verification of multi-model ensemble forecasts; the derived statistics can provide guidance for the design of future operational ensemble configurations without having to run additional ensemble forecast experiments for all the potential configurations.

Keywords: Multi-model ensemble, ensemble size, optimal weighting, predictive verification

1. Introduction

Ensemble systems provide a framework for probabilistic forecasting in numerical weather prediction. A collection of forecasts with the same target serves as a basis for the generation of probabilistic products. In this framework, it is well-established that the ensemble-size, that is the number of forecasts available at the product-generation stage, has an impact on the quality of the ensemble probabilistic products. This is for example the case when we consider a cumulative probability distribution function (CDF) generated from m ensemble members and the quality of the CDF forecasts estimated with the continuous ranked probability score (CRPS). When the ensemble is reliable, the ratio between the expected score of the m-member-based forecasts and the expected score if the ensemble was of infinite size is $1 + 1/m$ (Richardson, 2001).

More generally, ensemble-adjusted scores provide a means to estimate the ensemble-size effect on forecast performance assuming ensemble member exchangeability and stationarity of the error statistics. The concept of score adjustment allows one to derive an unbiased estimate of a score S for an ensemble of size M when say $m < M$ members are available for the score computation. Denoted $S_{m\to M}$, adjusted scores can be applied, for example, to compare ensemble forecasting systems with different ensemble-sizes, disentangling the ensemble-size effect and the impact of ensemble/model configuration on forecast performance. Furthermore, adjusted scores provide estimates of the expected benefit of an ensemble-size upgrade without the need to run extra members. Practically, in numerical experimentation, expected scores of an M-member ensemble are inferred from a small number of members (Leutbecher, 2018). This way, unused computational resources are made available for other experimental tests. Adjusted versions of the CRPS, Brier score, and ranked probability score are available in the

Corresponding author. e-mail: zied.benbouallegue@ecmwf.int

Citation: Tellus A: 2019, **72**, 1697165, https://doi.org/10.1080/16000870.2019.1697165

literature as well as an adjusted version of the ignorance score for forecasts issued as Normal distributions (Ferro et al., 2008; Siegert et al., 2018).

The first objective of this paper is to revisit the concept of ensemble-adjusted scores and its applicability in the general context of multi-model ensembles. The multi-model ensemble approach refers here to the combination of forecasts from $k (k>1)$ ensembles with different statistical characteristics. Let $m_1, \ldots, m_k$ denote the ensemble-size of the k ensembles that are going to be combined. An ensemble-adjusted score $S_{(m_1,\ldots,m_k)\to(M_1,\ldots,M_k)}$ provides a forecast performance estimate of a $M_1, \ldots, M_k$ combined ensemble forecast based on verification statistics from ensembles of size $m_1, \ldots, m_k$. In the following, we discuss how multi-model ensemble-size affects forecast performance, in particular in terms of the CRPS. As an application, the ensemble-size effect is investigated for a dual-resolution ensemble which combines forecasts from the same model but run at two different resolutions (Leutbecher and Ben Bouallègue, 2019).

The benefit of the multi-model ensemble approach and the rationale explaining its success were investigated successively in Hagedorn et al. (2005); Weigel et al. (2008); Weigel and Bowler (2009); Leutbecher and Ben Bouallègue (2019). Multi-model ensembles by definition gather ensemble forecasts with different error characteristics. Forecast improvement occurs as a result of a noise reduction associated with the increase of the ensemble-size or by addition of new predictable signals (DelSole et al., 2014). When the size of the combined ensemble is fixed, forecast improvement can arise from an appropriate weighting of the different ensemble members. Instead of a simple pooling of the forecasts, post-processing methods can be applied in order to attribute more weight to a set of forecasts when justified by previous forecast performance (see Doblas-Reyes et al., 2005; Casanova and Ahrens, 2009; DelSole et al., 2013; Baran et al., 2019, among others).

The second objective of this paper is to propose a new approach for ensemble-weighting optimisation. We show that optimal weights can be derived directly from the kernel representation of the CRPS. As a result, ensemble-size effect and weighting strategy can be analysed simultaneously. This is illustrated here in the particular case of a two-ensemble combination. An exhaustive analysis of weighted and unweighted ensemble combinations is performed without the need to run large ensemble experiments or complex post-processing methods. This novel approach to forecast verification is coined *predictive verification* which is used as an umbrella term for the assessment of potential ensemble configurations based on a reduced set of members. As a fundamental application, predictive verification of ensemble forecasts aims to provide guidance for the design of future ensemble systems.

This paper is organised as follows: the concept of ensemble-adjusted scores in a multi-model setting is described and tested in Section 2, applications to ensemble-size optimisation as well as model weighting are discussed in Section 3 before concluding in Section 4.

2. Score adjustment

2.1. Unbiased estimators

Consider forecasting a continuous outcome y. Consider first a single ensemble system of size m. The ensemble members are denoted $\mathbf{z}_m = (z_1, \ldots, z_m)$. For the derivation of score unbiased estimators, it is assumed that the ensemble members (from any one model) are exchangeable. We focus in this manuscript on the CRPS but unbiased estimators for the Brier score are also provided in Appendix B.

Following Gneiting and Raftery (2007), the kernel representation of the CRPS (C) for the empirical distribution function (EDF) of the ensemble reads:

$$C(\mathbf{z}_m, y) = \frac{1}{m}\sum_{j=1}^{m}|z_j - y| - \frac{1}{2m^2}\sum_{g=1}^{m}\sum_{h=1}^{m}|z_g - z_h| \qquad (1)$$

where the first term is the mean absolute error of the ensemble members and the second term is a measure of the ensemble spread. An unbiased estimator of the score for an ensemble with M members takes the following form:

$$C(\mathbf{z}_m, y) - \frac{1}{2}\frac{M-m}{M(m-1)m^2}\sum_{g=1}^{m}\sum_{h=1}^{m}|z_g - z_h|, \qquad (2)$$

as discussed in Ferro et al. (2008). The adjusted CRPS is denoted $C_{m\to M}$, and it is interpreted as the expected CRPS if we had M ensemble members, estimated from our statistical knowledge of the ensemble characteristics based on m ensemble members.

Now consider that we are in a multi-model setting. The multi-model ensemble comprises k models with m_i members from the i-th model. The multi-model ensemble forecast is denoted

$$\mathbf{z_m} = (z_{11} \quad \ldots \quad z_{ig} \quad \ldots \quad z_{km_k}) \qquad (3)$$

where z_{ig} is the g-th member in the i-th model. The multi-index $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ contains the ensemble sizes. We will refer to it as the (multi-model) ensemble size. The total ensemble size is $|\mathbf{m}| = \sum m_i$.

We can form the EDF for each of the k models and then combine these to form a probability forecast from the multi-model ensemble. We define the forecast distribution function to be the mixture

$$\sum_{i=1}^{k}\lambda_i F_i(z), \qquad (4)$$

where $\lambda_i>0$ is the weight assigned to the EDF, F_i, of model i and $\sum_{i=1}^{k}\lambda_i = 1$. The CRPS for this forecast distribution is

$$C(\mathbf{z_m},\lambda,y) = \sum_{i=1}^{k}\frac{\lambda_i}{m_i}\sum_{j=1}^{m_i}|z_{ij}-y| - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\lambda_i\lambda_j}{m_i m_j}\sum_{g=1}^{m_i}\sum_{h=1}^{m_j}|z_{ig}-z_{jh}|. \quad (5)$$

If we choose $\lambda_i = 1/k$ then each model receives equal weight, and if we choose $\lambda_i = m_i/\sum_j m_j$ then each member receives equal weight. We refer to this latter choice as *ensemble pooling* and we compare it to estimating *optimal weights* in Section 3.

Similarly to the single ensemble case, we would like to measure the expected ensemble-size effect on forecast performance in a multi-model ensemble setting. Not only is exchangeability of the ensemble members from any one model required but also *multi-model exchangeability*, which is a form of *partial exchangeability* (Bernardo and Smith, 2000), as well as multi-model ensemble size invariance. These so-called generalised multi-model exchangeability conditions do not require exchangeability between models. Formal definitions of the *multi-model exchangeability* and *multi-model ensemble size invariance* conditions are provided in Appendix A, Section A2 along with the corresponding mathematical developments. In addition, we provide more practical conditions that can be checked easily for any multi-model ensemble (see Equations (A7) and (A8)). In plain words, these conditions demand (i) that the expected mean absolute error of an ensemble member only depends on the model, i.e. it is the same for all members generated with that model and (ii) that the expected mean absolute difference between a pair of distinct members only depends on which models generated them, i.e. it is the same for all pairs of distinct members provided they originate from the same pair of models. Finally, the impact of the violation of the requirements is illustrated based on a concrete example in Section 2.3.

When the generalised multi-model exchangeability conditions are satisfied, an unbiased estimator of the CRPS for a multi-model ensemble with $\mathbf{M} = (M_1, \ldots, M_k)$ members is:

$$C(\mathbf{z_m},\lambda,y) - \frac{1}{2}\sum_{i=1}^{k}\lambda_i^2\frac{(M_i-m_i)}{M_i(m_i-1)m_i^2}\sum_{g=1}^{m_i}\sum_{h=1}^{m_i}|z_{ig}-z_{ih}|. \quad (6)$$

The adjusted CRPS in a multi-model setting is denoted $C_{\mathbf{m}\to\mathbf{M}}$. How the ensemble-adjusted CRPS can help with the design of a multi-model ensemble system is discussed later in Section 3. But first, in the remainder of this section, the robustness of Expression (6) as a score estimator is tested on a particular multi-model ensemble.

2.2. Multi-model ensemble setting

The concepts developed in Sections 2.1. (and later in Sections 3) are tested on a dual-resolution ensemble experiment. A dual-resolution ensemble is a particular case of a multi-model ensemble because the different contributing ensembles share the same underlying model. However, this specificity is neither required for the application of predictive verification as developed here, nor impacts the interpretation of the results. The choice of a multi-resolution ensemble to illustrate our method derives from a recent interest at ECMWF for this type of configuration but the method will also work with more traditional multi-model ensembles (as long as they satisfy the generalised multi-model exchangeability conditions).

In our test example, the forecast data-set comprises forecasts from the same numerical model (the ECMWF integrated forecasting system) but run at two different resolutions: TCo639 (∼18km) and TCo399 (∼29km). They are referred to as the higher resolution (Hres) and the lower resolution (Lres) members, respectively. A dual-resolution ensemble combines p Lres and q Hres members into a so-called (p,q) ensemble. The combination parameters, p and q, can be varied in such a way, for example, to keep constant the total computational cost of an ensemble run. The cost ratio between forecasts at TCo639 and TCo399 is approximately 4:1. Given a fixed computational cost defined by limited computer resources, one would like to know the combination (p,q) which optimises the predictive performance of the ensemble.

In order to answer this question, several forecast combinations with the same computational cost are assessed and their performances compared. An ensemble with an operational-like setup, that is a (0,50) ensemble which comprises 50 Hres members only, is used as the reference ensemble forecast. Other ensemble combinations are compared with this reference. These baseline combinations correspond to the (40,40), (120,20), (160,10), and (200,0) ensembles. Results of this type of analysis are documented in Leutbecher and Ben Bouallègue (2019). They show that the (40,40) ensemble performs significantly better than the other tested combinations when focussing on 2 m temperature short- and medium-range forecasts over the northern hemisphere.

Here, the question is whether the same results and conclusion can be obtained using ensemble-adjusted scores. The ensemble-adjusted approach potentially allows one to reduce drastically the cost of an ensemble experiment. In the case of a positive answer to the above question, score adjustment would provide the framework for the analysis of all potential ensemble combination performance at a lower experiment computational cost (see Section 3).

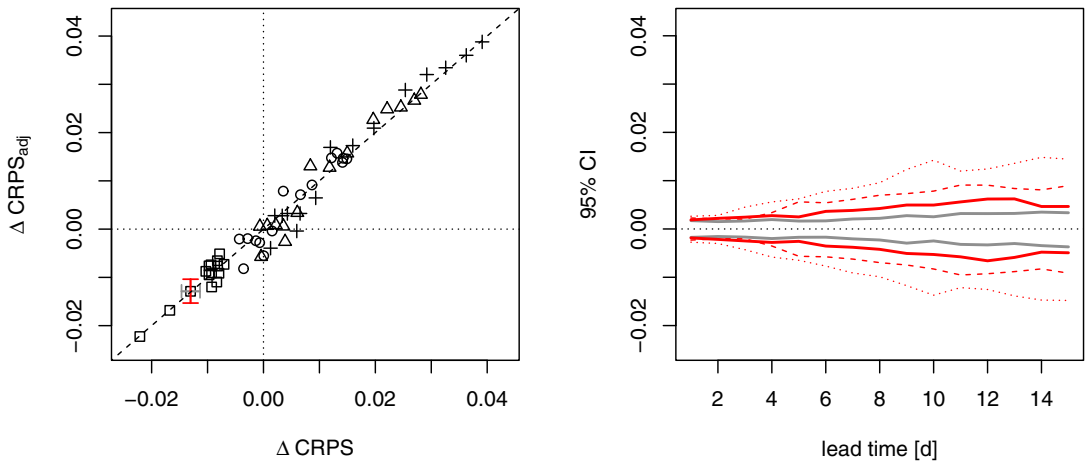

Fig. 1. **Left**: comparison of scores computed from actual (p, q) ensembles ($\Delta CRPS$) and adjusted scores based on a (8) subset of members ($\Delta CRPS_{\text{adj}}$). The scores plotted here are CRPS normalised differences between the (0,50) reference ensemble and the baseline multi-model combinations (40,40), (20,120), (10,160), and (200,0) as represented by squares, circles, triangles, and crosses, respectively. Each symbol corresponds to the result for one lead-time ranging between 1 and 15 days. **Right**: amplitude of the confidence intervals (CI) associated with the CRPS normalised differences (grey line) and the adjusted CRPS normalised differences (red lines) based on (2, 4), and (8) subsets of members (dotted, dashed, full lines, respectively), as a function of the forecast lead-time. CI are estimated by block-bootstrapping with blocks of three days. Results are shown for the comparison between the (0,50) and the (40,40) ensembles only. CI at day 3 based on a (8) subset of members are reported on the plot on the left. Note that the vertical axes have the same scale in both plots.

Using Expression (6), performance analysis of multi-model combinations is based on verification statistics computed from small subsets of Hres and Lres members. Subsets of the type (2, 4), and (8) are tested where each forecast of the subset is selected in order to be exchangeable with the other members. Leutbecher (2018) provides more details about member selection and exchangeability of the ECMWF ensemble.

Dual-resolution ensemble performance is assessed for several surface weather variables but only results for 2 m temperature forecasts are shown here. Results for 10 m wind speed and 24 h accumulated total precipitation were analysed as well but only briefly discussed here because they are qualitatively similar. The chosen verification period covers the boreal summer (JJA) 2016, and the forecasts are compared with SYNOP measurements distributed over the northern hemisphere.

2.3. Illustration

Performance in terms of CRPS is computed for each baseline combination, (40,40), (120,20), (160,10), and (200,0), and compared with the performance of the reference forecast (0,50). The CRPS difference is normalised by the mean CRPS of the reference forecast over the verification period and is simply denoted $\Delta CRPS$. A negative difference indicates that the baseline combination is better than the reference forecast. In terms of experiment computational cost, this analysis requires, in our example, to run 50 Hres members and 200 Lres members over a 92-day verification period.

Now, the CRPS for each of the baseline/reference combinations is estimated based on a (8) subset of members. In other words, we compute $C_{(8,8)\to(40,40)}$, $C_{(8,8)\to(120,20)}$, and so on using Expression (6). The score differences are then normalised by the mean scores of the reference combination ($C_{(0,8)\to(0,50)}$) and denoted $\Delta CRPS_{\text{adj}}$ because they are based on *adjusted* scores. This time, the ensemble performance analysis requires only 8 Hres members and 8 Lres members. This corresponds approximately to 10% of the experiment computational cost of the original analysis. Adjusted scores based on smaller subsets, (4) and (2), are also tested. They correspond to a reduction of the computational cost by a factor 20 and 40, respectively.

Figure 1 shows the correspondence between $\Delta CRPS$ and $\Delta CRPS_{adj}$ based on a (8) subset of members. Performance for 2 m temperature forecasts is shown for the four baseline combinations at each forecast lead-time (ranging between day 1 and day 15) separately (15×4 points in total). Very good agreement appears between normalised score differences computed from the actual dual-resolution ensembles and estimated from the subset of members. The corresponding correlation coefficient reaches 0.99 and the rank correlation coefficient is 0.88 as reported in Table 1. When the size of the ensemble subset on which the estimation is based is smaller than

Table 1. Gain in experiment computational time and accuracy of the score estimates based on adjusted scores for different sizes of ensemble subset.

(p,q) subset	(2)	(4)	(8)
relative computational time [%]	2.5	5	10
Kendal-τ corr. coeff.	0.73	0.88	0.88
Pearson corr. coeff.	0.94	0.98	0.99

(8), the accuracy of the score estimates tends to decrease. In all cases, the linear correlation between normalised score differences $\Delta CRPS$ and $\Delta CRPS_{\text{adj}}$ is higher than 0.9, even when based on the computationally very cheap (2) subset of ensembles. Performing the same analysis for 24 h precipitation and 10 m wind speed, we notice that the correlation coefficients for these variables are in general smaller than for 2 m temperature (not shown). However, the linear correlation coefficient reaches 0.87 for the former and 0.97 for the latter when based on a (8) ensemble. In principle, the method proposed here can be applied similarly to any other weather variable.

Computational resources for experimental testing are generally limited. In order to decrease computational cost for numerical experimentation, one can think of reducing the length of the testing period. This alternative is also considered here. Scores computed from the actual size dual-resolution ensembles but over reduced sets of randomly selected verification days are compared with scores averaged over a full 92-day verification period (JJA 2016). Following the same procedure as for the ensemble-adjusted scores, correlations between normalised score differences and their estimates are computed for a range of verification window lengths. The results are reported in Table 2. It appears that reducing the number of verification days can provide results substantially different than results obtained with the original verification sample. For example, a reduction in computational cost to ~10% of the original cost leads to a rank correlation coefficient below 0.8. Comparing results in Tables 1 and 2, we see that estimates from ensemble-adjusted scores are more robust than estimates based on a reduced sample of observations with comparable experiment computational time. This is the case not only for 2 m temperature forecasts but also for 24 h precipitation and for 10 m wind speed forecasts (not shown).

Besides the accuracy of the ensemble-adjusted scores, the level of confidence associated with score differences is also important when verification results serve as a basis for decision-making regarding future ensemble configurations. In Fig. 1 (right panel), we show as a function of the forecast lead-time the uncertainty associated with the score differences, $\Delta CRPS$ (grey lines) and $\Delta CRPS_{\text{adj}}$ based on a (8) ensemble (solid red lines), on a (4) ensemble (dashed red lines), and on a (2) ensemble (dotted red lines). Score uncertainty is represented by the 2.5%–97.5% percentile of the block-bootstrapped score distribution, using a block length of three days. In general, the score uncertainty tends to increase with forecast lead-time. The score uncertainty increases more rapidly when the performance is estimated with ensemble-adjusted scores. The uncertainty of the adjusted scores increases with decreasing size of the ensemble used for the score estimations as illustrated in Fig. 1 (right panel) for adjusted scores estimated from (4, 8) and (2) ensembles. The level of uncertainty (right panel) with respect to the score differences (left panel) appears however reasonable in this example. The CRPS difference $\Delta CRPS$ is significant at a 5% level in 85% of the cases (for all lead times and combinations). When score estimates are based on a (8) ensemble, the ratio of significant $\Delta CRPS_{adj}$ values is still 83%. For example, in Fig. 1 (left panel), the benefit of the (40,40) combination at day 3 is significant when computing both $\Delta CRPS$ and $\Delta CRPS_{\text{adj}}$ as depicted by the grey and red confidence bars, respectively. However, when the score estimations are based on (4) and (2) ensembles, the percentage of significant differences falls to 78% and 57%, respectively.

Table 2. Same as Table 1 but for score estimates based on different verification sample sizes and using the actual complete ensemble with p Lres and q Hres members.

Number of verification days	5	10	19
Relative computational time [%]	5.4	10.8	15.2
Kendal-τ corr. coeff.	0.68	0.76	0.82
Pearson corr. coeff.	0.90	0.93	0.97

Finally in this section, we would like to highlight the importance of the generalised multi-model exchangeability condition. This condition is required in order to have a valid unbiased score estimator as discussed in Section 2.1. For illustration purposes, we consider a configuration where the generalised exchangeability conditions are violated although the individual single ensembles are still exchangeable: In the following example members from the two models share the same initial perturbations. Every Hres member has the same initial condition as one corresponding Lres member. This implies that the difference between such a pair of Hres and Lres members sharing the initial condition is on average smaller than the difference between any other pair of Hres and Lres members. Figure 2 shows the adjusted score estimation for this example that violates exchangeability (bottom panel) as well as for a configuration that satisfies the generalised multi-model exchangeability condition (top panel). In the latter case, the score estimate is unbiased while in the former case, the CRPS estimate exhibits a clear bias up to day 6. For any actual ensemble generation methodology, it can be checked whether the exchangeability conditions (A7) and (A8) expressed in the Appendix hold within

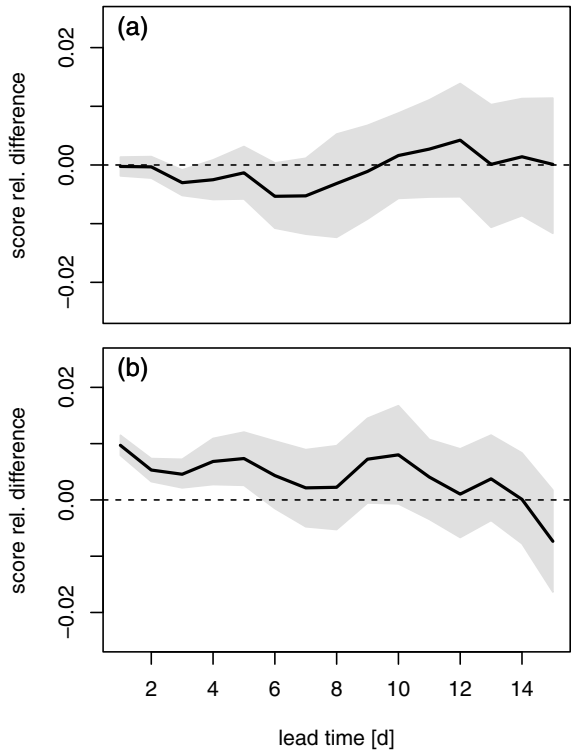

Fig. 2. Relative difference between CRPS and adjusted CRPS for a (40,40) ensemble as a function of the forecast lead time. The score adjustments are based on a (2) ensemble subset. Generalised multi-model exchangeability conditions are respected in (a) and violated in (b). Mean difference over the verification period (black curve) and variability as measured by block-bootstrap 90% confidence intervals (grey plume).

sampling uncertainty. Multi-model ensembles based on ensembles from very different models might of course fulfil these conditions, but it is worth noting that if any one of the models contributing to the multi-model ensemble fails to respect the criterion of member exchangeability, the approach proposed here is not applicable. Similarly, as in Leutbecher (2018), we can affirm that it will be more difficult to infer scores from a small subset when a multiphysics-based ensemble is contributing to the assessed multi-model ensemble because different members will have different biases. These are likely to render the expected mean absolute error dependent on the physics choices of the member and it is likely to render the expected mean difference of a pair of members dependent on the physics choices of this pair.

3. Multi-model ensemble design

In this section, the concept of ensemble-adjusted scores is exploited to efficiently assess the ensemble-size effect on multi-model ensemble performance. This is illustrated in the context of the design of a dual-resolution ensemble as described in Section 2.2. Here, again, the target is the minimisation of the ensemble expected CRPS. First, optimal mixtures of higher and lower resolution ensemble forecasts for a given computational cost are investigated in Section 3.1. Second, optimal weighting strategies for fixed combinations of members are discussed in Section 3.2. Both applications are complementary. The illustrations are based on the case where $k = 2$, that is when forecasts from two different models are combined, but the methodology can be generalised to the combination of any number of models.

3.1. Ensemble pooling

We first consider flexibility in terms of ensemble-size in the design of a multi-model ensemble. This is for instance the case in the dual-resolution ensemble example: A decision can be made about the number of Lres and Hres forecasts to be combined with computer power limited by current or foreseeable resources. Using ensemble-adjusted scores, forecast performance of any combination is estimated from a small set of Hres and Lres forecasts. These estimates of the scores for many different configurations are available virtually for free once the required verification statistics from one representative small ensemble have been computed. There is no additional cost in terms of numerical experimentation or in terms of repeated computations of verification statistics.

As suggested by results in Section 2.3, we illustrate the adjusted score applications by deriving statistics from an (8) ensemble. Indeed, in that case, ensemble-adjusted CRPS are expected to provide robust performance estimations for larger ensemble combinations. Adjusted CRPS are estimated here for a range of model combinations (p,q) with $p \in [0, 200]$ and $q \in [0, 50]$.

In Fig. 3, results are shown for 2-metre temperature forecast at day 5. On the left panel, a simple pooling of all combined members is considered. To ease readability, the CRPS is normalised by the minimum CRPS over all tested combinations: this shows the percentage of deterioration with respect to this (local) optimal score, indicated with an asterisk on the graph. Among all configurations considered here, the optimal one is not the (200,50) but the (25,50): adding more lower resolution members to the ensemble degrades the ensemble performance in this example. Configurations with equal predictive performance are highlighted with contour lines (solid black lines). Configurations with computational cost equivalent to the current one are indicated with a dashed descending diagonal line.

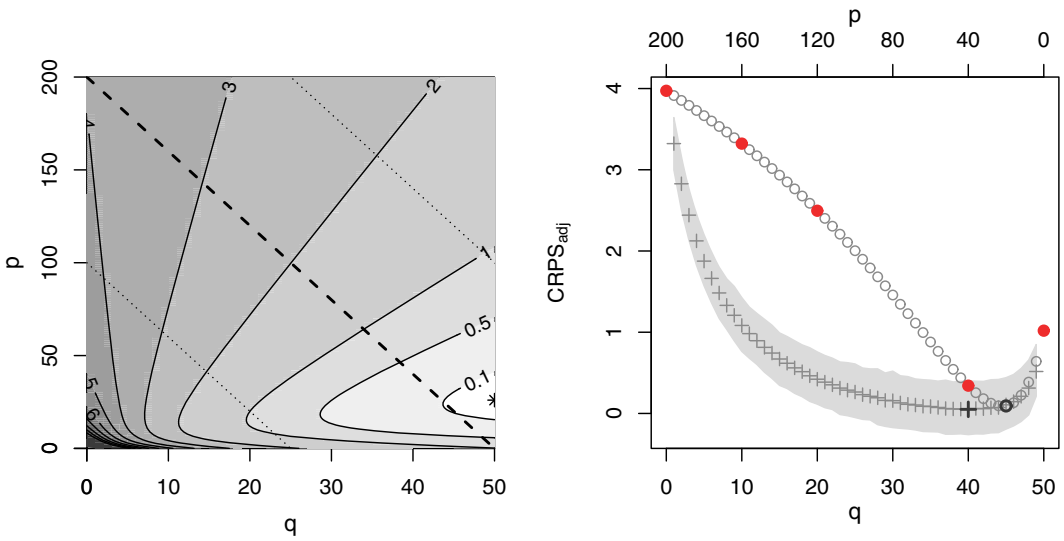

Fig. 3. **Left**: ensemble-adjusted CRPS of a dual-resolution ensemble as a function of the number of Lres (p) and Hres (q) members combined with equal weighting. The CRPS values are normalised in order to indicate the percentage degradation with respect to the optimal solution among the tested combinations (as indicated by a *). Black lines indicate ensemble combinations with equal performance. The diagonal dashed line indicates ensemble combinations with computational cost equivalent to the (0,50) reference forecast. Dotted lines indicate results for ensemble combinations with half or double the reference computer resources. **Right**: ensemble-adjusted CRPS ($CRPS_{adj}$) after normalisation as a function of the number of Hres (Lres) members q (p) considering a fixed computational cost equivalent to running 50 Hres forecasts. The plot shows performance for ensembles with equal weighting (○) and optimal weighting (+) for each combination. Weights are estimated based on a (8) ensemble. Grey shading indicates 90%-confidence intervals (see text). Optima are highlighted in bold. Results for the baseline/reference combinations of the original analysis are indicated in red. Results are valid for 2 m temperature forecast at day 5.

Parallel lines to the descending diagonal indicate results for combinations with equal computational costs. For example, results for combinations that require twice and half the current level of computer resources are indicated with dotted lines. Focussing on the current computational cost (dashed diagonal), one can consider running a (200,0) ensemble (top left corner), or a (0,50) ensemble (bottom right corner), or any intermediate combination. The ensemble performances of all these ensemble combinations with equal computational cost are reported on the right panel in Fig. 3.

In Fig. 3 (right panel), the ensemble-adjusted CRPS provides estimates of the dual-resolution ensemble performance for the full range of possible combinations between Lres and Hres members given the current computer resources (grey circles). The original analysis focussed only on four baseline plus one reference combinations (red circles) pointing to a (40,40) ensemble as the optimal combination. Applying score adjustments, a finer analysis shows that the (20,45) ensemble is the optimal dual-resolution combination for this particular weather variable and forecast lead time (2 m temperature at day 5). Using the adjusted CRPS in Expression (6), this type of analysis can be repeated easily, and at no additional cost for any other computational resource constraint.

3.2. Ensemble weighting

We consider now the case where the number of forecasts to be combined is fixed. The question is whether the ensemble performance can be improved by applying appropriate weighting to each combined model. We propose here an analytical expression of the optimal weights which is directly derived from the kernel representation of the CRPS. In contrast to post-processing methods generally in use, no numerical optimisation procedure is required in the proposed approach. Nevertheless, prior knowledge of forecast performance based on historical data is needed.

The weighting strategy proposed here is useful for showing the relative merits of the different model ensembles and its expected impact on multi-model ensemble performance. Potentially, further improvement could be achieved with considering additionally bias correction of the ensemble members. Such corrections are left for future studies.

In the following, we discuss the case where forecasts from two models are directly combined ($k = 2$). We use the simplified notations defined in Appendix A, Section A1. Mathematical developments can be found in Appendix A, Section A4 where a general solution for

optimal weighting is presented for an arbitrary number $k \geq 2$ of models.

Considering the constraint $\lambda_1 + \lambda_2 = 1$, the minimisation problem follows:

$$\underset{\lambda_1}{\operatorname{argmin}}(C_{(m_1,m_2)}(\lambda_1)). \tag{7}$$

The optimal weighting depends directly on the number of members combined from model 1 and model 2, that is m_1 and m_2, respectively. Optimal weights can also be estimated accounting for the ensemble-size effect. In order to derive optimal weights for a (M_1, M_2) ensemble based on a subset (m_1, m_2) of forecasts, one must solve

$$\frac{dC_{(m_1,m_2)\to(M_1,M_2)}(\lambda_1)}{d\lambda_1} = 0. \tag{8}$$

Using the kernel representation of the adjusted CRPS in a multi-model context, an analytical solution of Equation (8) can be found applying linear algebra. The optimal weight estimate follows:

$$\hat{\lambda}_1^{\circ} = \frac{\hat{C}_2 - \hat{C}_1 + \hat{R}_{12}}{2\hat{R}_{12}}, \tag{9}$$

where the first two terms on the numerator correspond to the adjusted CRPS of model 2 and model 1, respectively, and with $\hat{R}_{12}$ defined as:

$$\hat{R}_{12} = 2\hat{D}_{12} - \hat{D}_{11} - \hat{D}_{22} \tag{10}$$

where $\hat{D}_{ij}$ is an estimate of the expected absolute difference between members of model i and members of model j. So $\hat{R}_{12}$ is proportional to the difference between 'inter-model spread' and mean 'intra-model' spread. This term is positive as soon as there is less similarity between members originating from two different models than between members originating from the same model.

From Equation (9) and the developments in Appendix A, Section A4, we can make the following remarks regarding the values that $\hat{\lambda}_1^{\circ}$ can take:

- $\hat{\lambda}_1^{\circ} = 0.5$ if $\hat{C}_1 = \hat{C}_2$, that is model 1 and 2 have the same performance;
- $\hat{\lambda}_1^{\circ} = 1$ if $\hat{C}_1 = 0$ and $\hat{C}_2 \neq 0$, that is model 1 is perfect and model 2 is not;
- $\hat{\lambda}_1^{\circ} = 0$ if $\hat{C}_2 = 0$ and $\hat{C}_1 \neq 0$, that is model 2 is perfect and model 1 is not.

Exploiting these results, performance of dual-resolution ensembles can be now examined considering optimal weighting. For each (p, q) ensemble, optimal weights are estimated using Equation (9) and applied to the score estimation in Expression (6). For illustration purposes, we apply an out-of-sample approach for the computation of the weights and their application to the multi-ensemble forecast before verification. We proceed as follow: the 92-day verification sample is divided into a training period and a testing period. Elements of the training group are randomly selected as 15 blocks of three consecutive days with replacement, that is blocks that can overlap. Days not selected for training are used for testing. We average the CRPS over all pairs in the training period and then compute the optimal weights for the mean CRPS. Optimal weights estimated over the training period are applied to the forecasts over the testing period. Testing and training period are then swapped. Verification of the weighted ensemble forecasts is finally computed over the whole verification window. The process is repeated 1000 times with a random selection of training and testing days in order to obtain a score distribution from which statistics are derived.

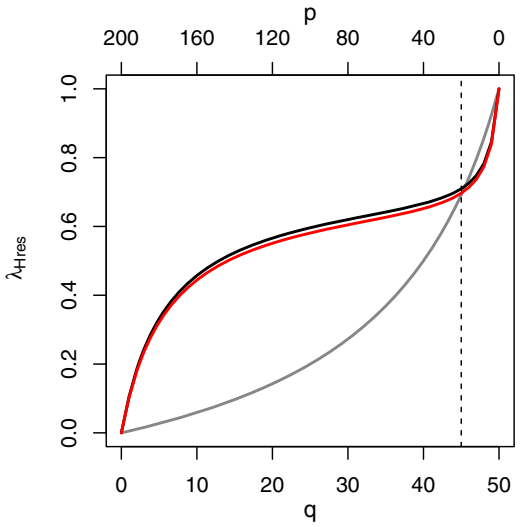

Fig. 4. Weight associated with the Hres model as a function of the number of Hres (Lres) members q (p) considering a fixed computational cost equivalent to running 50 Hres forecasts: weights when ensemble pooling is applied (grey line), optimal weights estimated from a (200,50) ensemble (black line), optimal weights estimated from a (8) ensemble (red line). The vertical dotted line indicates the optimal (p,q) combination as seen in Fig. 3 (right panel).

Figure 4 shows the optimal weights as estimated from a (200,50) ensemble (black line) and from a (8) subset of members (red line). Optimal weighting is compared with ensemble pooling where all combined members have the same weight (grey line). When a coloured line is above the grey line, it means that the weight optimisation increases the contribution of the Hres members with respect to the ensemble pooling. This is the case when the number of combined Hres members is smaller than the one associated with the optimal (p, q) combination as indicated by the vertical line. We also see that the differences between optimal weight estimates based on a (200,50) ensemble or a (8) ensemble are small. Moreover, the resulting CRPS when applying one or the other

optimal weight exhibit differences which are not larger than 0.1% and non-significant (not shown).

In Fig. 3 (right panel), ensemble forecast performances with optimal weighting are plotted for each ensemble combination (p,q). The CRPS of the forecast with optimal weights (crosses) can be compared with the CRPS of pooled forecasts (circles) for a given combination of Hres and Lres forecasts. Grey shading shows confidence intervals as measured by the 5%–95% percentiles of the score distribution. They represent the performance uncertainty associated with the weighting procedure.

From the results presented on the right panel in Fig. 3, we conclude that (i) the application of optimal weights can substantially improve the performance of a multi-ensemble forecast, (ii) a large range of ensemble combinations have near-optimal score when optimal weighting is applied, and (iii) the optimal combination with pooled forecasts (the (20,45) ensemble) is not improved further by model weighting.

More generally, the multi-model ensemble performance under simple pooling and optimal weighting provide complementary information. For the design of an ensemble system, the assessments of raw and of potential performance after optimal weighting are both relevant figures. Both can be performed efficiently based on the ensemble-adjusted CRPS.

4. Summary

Ensemble-adjusted scores allow one to account for the ensemble-size effect on ensemble forecast performance. This paper revisits the ensemble-adjusted score concept in the context of multi-model ensemble forecasting. An unbiased estimator of the continuous ranked probability score as a function of the ensemble size is proposed and its robustness tested on dual-resolution ensemble forecasts. It is shown that adjusted scores $S_{(m_1,\ldots,m_k)\to(M_1,\ldots,M_k)}$ based on a small subset of exchangeable members from each model (typically $m_i = 8$ for any combined models i) provide good performance estimates of any $(M_1, \ldots, M_k)$ ensemble configuration. The validity of the approach depends on generalised multi-model exchangeability conditions. A simple tool is provided to determine whether the multi-model ensemble at hand satisfies the conditions within sampling uncertainty. Score adjustment of the Brier score in a multi-model ensemble context is also provided in the appendix of this article. Further research is welcome in order to develop ensemble-size adjustment of other scores for a more comprehensive analysis of ensemble performance based on small subsets of ensemble members.

From a research testing perspective, the use of ensemble-size adjusted scores can represent a substantial saving in terms of the computational cost for the numerical experimentation. In our illustrative example, a decrease up to a factor 10 of the experiment cost (by running fewer members) does not considerably deteriorate the quality of the analysis: The unbiased score estimates are highly correlated with scores computed from the actual size ensemble. It is shown that this strategy is more efficient than a strategy consisting in drastically reducing the verification sample in terms of the number of forecast start dates. The latter can be detrimental to a robust assessment of ensemble performance.

Ensemble-adjusted scores find applications in the design of multi-model ensemble systems. This is also illustrated here with a dual-resolution ensemble where an optimal combination of higher- and lower-resolution forecasts is targeted. Not only simple pooling of forecasts but also optimal weighting of the contributing models can be investigated, accounting for the ensemble-size effect. Based on linear algebra, optimal weights are directly derived from the CRPS kernel representation. Applying optimal weighting strategies helps to demonstrate the potential performance of optimally combined ensemble forecasts. The derivation of optimal weights, in a non-iterative fashion, can be applied without restriction to any combination of ensemble members.

At low experiment computational cost and with limited verification effort, it is possible to draw a full picture of expected performance in terms of CRPS as a function of the number of members from each contributing model. The optimal ensemble configuration can be easily identified for a given computational cost with and without weighting members. This new type of ensemble skill investigations is coined predictive verification and aims to provide a framework for making informed decisions on future multi-model ensemble configurations.

Acknowledgement

The authors are grateful to Francisco J. Doblas-Reyes for triggering their attention on the main topic of this manuscript during the Annual Seminar held at ECMWF in September 2017.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Baran, S., Leutbecher, M., Szabó, M. and Ben Bouallègue, Z. 2019. Statistical post-processing of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.* **145**, 1705–1720.

Bernardo, J. M. and Smith, A. 2000. *Bayesian Theory*. Wiley, Chichester.

Casanova, S. and Ahrens, B. 2009. On the weighting of multimodel ensembles in seasonal and shortrange weather forecasting. *Mon. Wea. Rev.* **137**, 3811–3822. doi:10.1175/2009MWR2893.1

DelSole, T., Nattala, J. and Tippett, M. 2014. Skill improvement from increased ensemble size and model diversity. *Geophys. Res. Lett.* **41**, 7331–7342. doi:10.1002/2014GL060133

DelSole, T., Yang, X. and Tippett, M. 2013. Is unequal weighting significantly better than equal weighting for multimodel forecasting? *Q. J. Meteorol. Soc.* **139**, 176–183. doi:10.1002/qj.1961

Doblas-Reyes, F., Hagedorn, R. and Palmer, T. 2005. The rationale behind the success of multimodel ensembles in seasonal forecasting I. Calibration and combination. *Tellus A* **57**, 234–252.

Ferro, C. A. T., Richardson, D. S. and Weigel, A. P. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Met. Apps.* **15**, 19–24. doi:10.1002/met.45

Gneiting, T. and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378. doi:10.1198/016214506000001437

Hagedorn, R., Doblas-Reyes, F. and Palmer, T. 2005. The rationale behind the success of multimodel ensembles in seasonal forecasting. I. Basic concept. *Tellus A* **57**, 219–233.

Leutbecher, M. 2019. Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.* **145(Suppl. 1)**, 107–128.

Leutbecher, M. and Ben Bouallègue, Z. 2019. On the probabilistic skill of dual-resolution ensemble forecasts. *Q. J. R. Meteorolog. Soc.* doi:10.1002/qj.3704

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

Siegert, S., Ferro, C. A. T., Stephenson, D. B. and Leutbecher, M. 2019. The ensemble-adjusted Ignorance score for forecasts issued as Normal distributions. Q. J. R. Meteorol. Soc. ***145(Suppl. 1)**, 129–139.*

Weigel, A. and Bowler, N. 2009. Comment on 'can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. Meteorol. Soc.* **135**, 535–539. doi:10.1002/qj.381

Weigel, A., Liniger, M. and Appenzeller, C. 2008. Can multi-model combination really enhance the prediction skill of ensemble forecasts?. *Q. J. Meteorol. Soc.* **134**, 241–260. doi:10.1002/qj.210

APPENDIX A: Mathematical details for CRPS adjustment

In the following, the mathematical details that were omitted from the main text are provided. The required exchangeability conditions and the proof that the score estimator given these conditions is unbiased are provided first. Then, the general solution for the optimal weights is derived.

A1. Kernel representation in compact notation

It is convenient to introduce a compact notation for the derivations that follow. The verification statistics that need to be aggregated in order to apply the score adjustment in the kernel representation of the CRPS are the mean absolute error E_i and the L1-spread matrix D_{ij}. When obtained from numerical experimentation with ensemble size $\mathbf{m}$, these verification statistics are

$$E_i(\mathbf{m}) = \frac{1}{m_i}\sum_{g=1}^{m_i}|z_{ig}-y|, \tag{A1}$$

$$D_{ij}(\mathbf{m}) = \frac{1}{2m_i m_j}\sum_{g=1}^{m_i}\sum_{h=1}^{m_j}|z_{ig}-z_{jh}|. \tag{A2}$$

The matrix $\mathbf{D}$ is symmetric. Its off-diagonal terms describe the diversity between models, or 'inter-model spread', while the diagonal terms describe the spread within the individual models, or 'intra-model' spread.

The CRPS kernel representation for the multi-model ensemble with weights λ as given earlier in Equation (5), can be expressed in terms of E_i and D_{ij} as

$$C_{\mathbf{m}}(\lambda) = \sum_{i=1}^{k}\lambda_i E_i(\mathbf{m}) - \sum_{i=1}^{k}\sum_{j=1}^{k}\lambda_i\lambda_j D_{ij}(\mathbf{m}). \tag{A3}$$

Similarly, the adjusted score according to (6) can be written as

$$\begin{aligned} C_{\mathbf{m}\to\mathbf{M}}(\lambda) &= C_{\mathbf{m}}(\lambda) - \sum_{i=1}^{k}\lambda_i^2\gamma_i D_{ii}(\mathbf{m}) \\ &= \sum_{i=1}^{k}\lambda_i E_i(\mathbf{m}) - \sum_{i=1}^{k}\sum_{j=1}^{k}\lambda_i\lambda_j\hat{D}_{ij}(\mathbf{m},\mathbf{M}), \end{aligned} \tag{A4}$$

with the adjusted spread matrix given by

$$\hat{D}_{ij} = D_{ij} + \delta_{ij}\gamma_i D_{ii} \tag{A5}$$

and ensemble-size adjustment factors

$$\gamma_i = \frac{(M_i - m_i)}{M_i(m_i-1)} \tag{A6}$$

where $\delta_{ij} = 1$ if $i=j$ and 0 otherwise. The $k + k(k+1)/2$ numbers E_i, D_{ij} with $i \le j$ characterise the joint distribution of y and $\mathbf{z_m}$ completely in terms of the CRPS. Storing the E_i and D_{ij} permits the computation of the CRPS estimator for any set of weights λ and any target ensemble size $\mathbf{M}$ from numerical experimentation with ensemble size $\mathbf{m}$.

A2. Generalised multi-model exchangeability conditions

Now, we focus on the conditions required to render the adjusted score given by Expressions (6), (A4) an unbiased estimate for the target ensemble size $\mathbf{M}$. The multi-model exchangeability and the ensemble size invariance conditions described in Sections A2.1 and

A2.2 are sufficient but impractical to validate. Subsequently in Section A2.3, we provide less general conditions in terms of expected absolute differences between members and between members and the verification data. The latter conditions are more practical as they can be checked easily for any multi-model ensemble.

A2.1. Multi-model exchangeability

We extend the notion of exchangeability to the multi-model setting as follows. For each of the k models consider an arbitrary permutation of its members. This describes a permutation for the entire ensemble that respects the order of the models in the vector $\mathbf{z_m}$. We can represent this permutation via its block-diagonal permutation matrix $\mathbf{P}$.

An ensemble composed of members from k different models is said to be multi-model exchangeable if for any such permutation $\mathbf{P}$ that consists of arbitrary permutations of the single-model sub-ensembles, the joint distribution of $\mathbf{Pz}$ is identical to the joint distribution of $\mathbf{z}$.

A2.2. Multi-model ensemble-size invariance

Consider an ensemble generation method that can generate multi-model exchangeable ensembles of different sizes, say $\mathbf{m} = (m_1, \ldots, m_k)$ and $\tilde{\mathbf{m}} = (\tilde{m}_1, \ldots, \tilde{m}_k)$. Let $\mathbf{z_m}$ and $\tilde{\mathbf{z}}_{\tilde{\mathbf{m}}}$ denote the two ensembles corresponding to the ensemble size $\mathbf{m}$ and $\tilde{\mathbf{m}}$. One can compare marginal distributions constructed from the two ensembles that contain the same number $l_j \leq \min(m_j, \tilde{m}_j)$ of distinct members from each model j. Let $\mathbf{H}$ and $\tilde{\mathbf{H}}$ denote operators that extract for each model j a subset of l_j members from the ensembles $\mathbf{z_m}$ and $\tilde{\mathbf{z}}_{\tilde{\mathbf{m}}}$, respectively. We define the multi-model ensemble generation method to be ensemble-size invariant if for any ensemble sizes $\mathbf{m}$ and $\tilde{\mathbf{m}}$ and for any subensemble extractions $\mathbf{H}$ and $\tilde{\mathbf{H}}$ that yield the same number of distinct members from each model j, the joint distribution of $\mathbf{Hz_m}$ is identical to the joint distribution of $\tilde{\mathbf{H}}\tilde{\mathbf{z}}_{\tilde{\mathbf{m}}}$.

A2.3. Conditions in terms of expected distances

For an ensemble that satisfies the generalised multi-model exchangeability conditions given in A2.1 and A2.2, the expected values of the distance between members and the distance between members and verification depend only on the model indices i, j and neither on ensemble size $\mathbf{m}$ nor on member numbers g, h within a model:

$$\mathbb{E}|z_{ig} - z_{jh}| = (1-\delta_{ij}\delta_{gh})\Delta_{ij}, \tag{A7}$$

$$\mathbb{E}|z_{ig} - y| = \Theta_i. \tag{A8}$$

These conditions are the key ingredient for the following proof.

A3. Proof

With the conditions expressed in Equations (A7) and (A8), the expected $\mathbf{E}$ and $\mathbf{D}$ are given by

$$\mathbb{E}\ E_i(\mathbf{m}) = \Theta_i, \tag{A9}$$

$$\mathbb{E}\ D_{ij}(\mathbf{m}) = \frac{1}{2}\Delta_{ij} \quad \text{for } i \neq j, \tag{A10}$$

$$\mathbb{E}\ D_{ii}(\mathbf{m}) = \frac{(m_i-1)}{2m_i}\Delta_{ii}. \tag{A11}$$

This implies that the expected adjusted spread matrix satisfies

$$\mathbb{E}\hat{D}_{ii}(\mathbf{m}, \mathbf{M}) = \frac{(M_i-1)}{2M_i}\Delta_{ii}. \tag{A12}$$

Therefore, $\mathbb{E}\hat{D}_{ii}(\mathbf{m}, \mathbf{M}) = \mathbb{E}D_{ii}(\mathbf{M})$ which yields the desired result that

$$\mathbb{E}\ C_{\mathbf{m}\to\mathbf{M}}(\lambda) = \mathbb{E}\ C_{\mathbf{M}}(\lambda). \tag{A13}$$

So $C_{\mathbf{m}\to\mathbf{M}}(\lambda)$ is an unbiased estimator for the expected CRPS at the target ensemble size, $\mathbf{M}$.

A4. Optimal weighting

Now, we describe how to obtain the optimal weights λ for a target ensemble size $\mathbf{M}$. Let

$$\mathbf{e} = (E_1 \quad \ldots \quad E_k)^{\mathrm{T}} \tag{A14}$$

denote the column vector with the mean absolute errors. Then, the adjusted CRPS can be written in matrix notation as

$$C_{\mathbf{m}\to\mathbf{M}}(\lambda) = \lambda^{\mathrm{T}}\mathbf{e} - \lambda^{\mathrm{T}}\hat{\mathbf{D}}\lambda. \tag{A15}$$

Optimal weights are sought subject to the constraint $\sum\lambda_j = 1$. This can be achieved via a Lagrange multiplier. Define

$$\mathcal{L}(\lambda, \phi) = C_{\mathbf{m}\to\mathbf{M}}(\lambda) - \phi(\mathbf{u}^{\mathrm{T}}\lambda - 1) \tag{A16}$$

with the vector $\mathbf{u} = (1 \quad \ldots \quad 1)^{\mathrm{T}}$. The optimum weights are then the solution of

$$\nabla_{\lambda,\phi}\mathcal{L} = 0. \tag{A17}$$

This yields linear equations for the weights and the Lagrange multiplier ϕ:

$$2\hat{\mathbf{D}}\lambda = \mathbf{e} - \phi\mathbf{u}, \tag{A18}$$

$$\mathbf{u}^{\mathrm{T}}\lambda = 1. \tag{A19}$$

Solving (A18) for the weights and inserting in (A19) yields the Lagrange multiplier

$$\phi = \frac{\mathbf{u}^{\mathrm{T}}\hat{\mathbf{D}}^{-}1\mathbf{e} - 2}{\mathbf{u}^{\mathrm{T}}\hat{\mathbf{D}}^{-}1\mathbf{u}}. \tag{A20}$$

Now, the optimum weights can be computed as

$$\hat{\lambda}^{\circ} = \frac{1}{2}\hat{\mathbf{D}}^{-}1\left(\mathbf{e} - \frac{\mathbf{u}^{\mathrm{T}}\hat{\mathbf{D}}^{-}1\mathbf{e} - 2}{\mathbf{u}^{\mathrm{T}}\hat{\mathbf{D}}^{-}1\mathbf{u}}\mathbf{u}\right). \tag{A21}$$

If we consider the combination of only two models ($k = 2$), the optimal weight associated with model 1 can be written as

$$\hat{\lambda}_1^{\circ} = \frac{\hat{C}_2 - \hat{C}_1 + \hat{R}_{12}}{2\hat{R}_{12}}, \tag{A22}$$

with $\hat{C}_i$ the expected value of the adjusted CRPS of model i in the single ensemble case,

$$\hat{C}_i = E_i - \hat{D}_{ii}, \tag{A23}$$

and with

$$\hat{R}_{12} = 2\hat{D}_{12} - \hat{D}_{11} - \hat{D}_{22}. \tag{A24}$$

From Equation (A22), it is straightforward to see that $\hat{\lambda}_1^{\circ} = 0.5$ when $\hat{C}_2 = \hat{C}_1$. We can also note that $\hat{\lambda}_1^{\circ} = 1$ when ensemble 1 is perfect (and ensemble 2 is not), that is $\hat{C}_1 = 0$ (and $\hat{C}_2 \neq 0$). Indeed, in that case, $z_{1g} = y$ with $g = 1, \ldots, m_1$ which implies that $\hat{D}_{11} = 0$ and $\hat{D}_{12} = \frac{1}{2}E_2$, so $\hat{R}_{12} = \hat{C}_2$. Similarly, we can show that $\hat{\lambda}_1^{\circ} = 0$ when ensemble 2 is perfect and ensemble 1 is not.

APPENDIX B. Adjusted brier score for multi-model ensembles

Consider forecasting a binary outcome, y. Suppose that we have $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ ensemble members from k models, and let N_i denote the number of the m_i members that forecast the event $\{y = 1\}$. Suppose that we form the probability forecast

$$P(\mathbf{m}, \lambda) = \sum_{i=1}^{k} \lambda_i \frac{N_i}{m_i}, \tag{B1}$$

a linear combination of the proportions, N_i/m_i, from each model with weights $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ associated with the k models. The Brier score for this probability forecast is

$$B(\mathbf{m}, \lambda) = \{P(\mathbf{m}, \lambda) - y\}^2. \tag{B2}$$

Suppose that we want to estimate $B(\mathbf{M}, \lambda)$, the Brier score that we would obtain if we had $\mathbf{M} = (M_1, M_2, \ldots, M_k)$ ensemble members from the k models. An unbiased estimate for the expected value of $B(\mathbf{M}, \lambda)$ is

$$B(\mathbf{m}, \lambda) - \sum_{i=1}^{k} \lambda_i^2 \frac{(M_i - m_i)}{M_i(m_i - 1)} \frac{N_i}{m_i} \left(1 - \frac{N_i}{m_i}\right). \tag{B3}$$

This result holds under the same generalised multi-model exchangeability conditions as the CRPS result (see Appendix A, Section A2).