The Generalized Discrimination Score for Ensemble Forecasts

ANDREAS P. WEIGEL

Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

SIMON J. MASON

International Research Institute for Climate and Society, Columbia University, Palisades, New York

(Manuscript received 5 November 2010, in final form 25 February 2011)

ABSTRACT

This article refers to the study of Mason and Weigel, where the generalized discrimination score D has been introduced. This score quantifies whether a set of observed outcomes can be correctly discriminated by the corresponding forecasts (i.e., it is a measure of the skill attribute of discrimination). Because of its generic definition, D can be adapted to essentially all relevant verification contexts, ranging from simple yes–no forecasts of binary outcomes to probabilistic forecasts of continuous variables. For most of these cases, Mason and Weigel have derived expressions for D , many of which have turned out to be equivalent to scores that are already known under different names. However, no guidance was provided on how to calculate D for ensemble forecasts. This gap is aggravated by the fact that there are currently very few measures of forecast quality that could be directly applied to ensemble forecasts without requiring that probabilities be derived from the ensemble members prior to verification. This study seeks to close this gap. A definition is proposed of how ensemble forecasts can be ranked; the ranks of the ensemble forecasts can then be used as a basis for attempting to discriminate between corresponding observations. Given this definition, formulations of D are derived that are directly applicable to ensemble forecasts.

1. Introduction

It is a well-established fact that the quality of a set of forecasts cannot be adequately summarized by a single metric, but requires that several attributes of prediction skill are considered (Murphy 1991). A fundamental skill attribute is “discrimination” (Murphy and Winkler 1992). Discrimination measures whether forecasts differ when their corresponding observations differ. For example, do forecasts for days that were wet indicate more (or less) rainfall than forecasts for days that were dry? If on average the forecasts indicate about the same amount of rainfall regardless of how much rain is actually received, then the forecasts are unable to discriminate wetter from drier days. Even a perfectly calibrated forecast system is effectively useless if it lacks discriminative power.

Recently, Mason and Weigel (2009, hereafter MW09) introduced the “generalized discrimination score” D , a generic verification framework that measures discrimination

and is applicable to most types of forecast and observation data. MW09 have derived formulations of D for observation data that are binary (e.g., “precipitation” vs “no precipitation”), categorical (e.g., temperature in lower, middle, or upper tercile), or continuous (e.g., temperature measured in °C); and for forecast data that are binary, categorical, continuous, discrete probabilistic (e.g., probability for temperature being in upper tercile), or continuous probabilistic (e.g., continuous probability distribution for temperature in °C). However, no guidance has been provided on how to calculate D for ensemble forecasts. It is the aim of this study to fill this gap.

MW09 have provided an in-depth discussion of the properties of D . One of the most appealing properties is the simple and intuitive interpretation of D : the score measures the probability that any two (distinguishable) observations can be correctly discriminated by the corresponding forecasts. Thus, D can be interpreted as an indication of how often the forecasts are “correct,” regardless of whether forecasts are binary, categorical, continuous, or probabilistic. For a given set of forecast–observation pairs, D is calculated as illustrated in Fig. 1. First, all possible (and distinguishable) sets of two

Corresponding author address: Andreas Weigel, MeteoSwiss, Kräbühlstrasse 58, P.O. Box 514, CH-8044 Zürich, Switzerland.
E-mail: andreas.weigel@meteoswiss.ch

FIG. 1. Conceptual illustration of the generalized discrimination score D . First, all possible (and distinguishable) sets of two forecast–observation pairs are constructed from the verification data. Then, for each of these sets, the question is asked whether the forecasts can be used to correctly rank the observations. Here D is given by the proportion of successful rankings.

forecast–observation pairs are constructed from the verification data. Then, for each of these sets, the question is asked whether the forecasts can be used to successfully distinguish (i.e., rank) the observations. The proportion of sets where this is the case yields the generalized discrimination score D . If the forecasts do not contain any useful information, then the probability that the forecasts correctly discriminate two observations is equivalent to random guessing (viz., 50%) and one would obtain $D = 0.5$. The more successfully the forecasts are able to discriminate the observations, the closer the score is to 1. On the other hand, forecasts that consistently rank the observations in the wrong way, would yield $D = 0$. For some data types, D is equivalent or similar to tests and scores that are already widely used in forecast verification and known under different names. For instance, if binary forecasts and observations are considered, D is a transformed version of the true skill statistic, also known as Pierce’s skill score (Pierce 1884). If forecasts and observations are measured on a continuous scale, D is a transformed version of Kendall’s ranked correlation coefficient τ (Sheshkin 2007). And if the forecasts are issued as discrete probabilities of binary outcomes, D is equivalent to the trapezoidal area under the relative operating characteristic (ROC) curve and to a transformation of the Mann–Whitney U statistic (Mason and Graham 2002).

How can D be calculated for ensemble forecasts? Despite their probabilistic motivation, ensemble forecasts are a priori not probabilistic forecasts, but “only” finite sets of deterministic forecast realizations. To derive probabilistic forecasts from the ensemble members, further assumptions concerning their statistical properties are required (Bröcker and Smith 2008). The question as to how D can be calculated therefore depends on how the ensembles are interpreted (i.e., whether they are seen as finite samples from underlying forecast distributions, or whether they have been converted into probabilistic forecasts). In the latter case, probabilistic versions of D , such as the area under the ROC curve, can be applied as described in MW09. However, D then inevitably not only measures the quality of the prediction system, but also the appropriateness of the probabilistic interpretation applied. In section 2, we show how D can be *directly* calculated for “raw” ensemble forecasts without requiring that probability forecasts are derived first. These formulations are illustrated with examples in section 3, and conclusions are given in section 4.

2. The discrimination score for ensemble forecasts

The calculation of D requires a definition of how to discriminate, or essentially rank, two ensemble forecasts.

If forecasts are issued as deterministic forecasts on a continuous scale, it is trivial to decide which one of two forecasts $\mathbf{y}_1$ and $\mathbf{y}_2$ is larger and should therefore (if the forecasts are skillful) indicate the larger one of the two corresponding observations. This decision is less obvious for ensemble forecasts. Consider for instance 3 hypothetical 5-member ensemble forecasts of temperature ($^{\circ}\text{C}$) with $\mathbf{y}_1 = (22, 23, 26, 27, 32)$, $\mathbf{y}_2 = (28, 31, 33, 34, 36)$, and $\mathbf{y}_3 = (24, 25, 26, 27, 28)$. While most people would intuitively label $\mathbf{y}_2$ larger than $\mathbf{y}_1$ and $\mathbf{y}_3$, the situation is less obvious when comparing $\mathbf{y}_1$ and $\mathbf{y}_3$. We therefore start by introducing a definition of how to rank ensembles, and based on that then derive a formulation of D for ensemble forecasts.

a. Ranking ensemble forecasts

Consider two ensemble forecasts $\mathbf{y}_s = (y_{s,1}, \dots, y_{s,m_s})$ and $\mathbf{y}_t = (y_{t,1}, \dots, y_{t,m_t})$, with m_i being the number of ensemble members of forecast $\mathbf{y}_i$, and $y_{i,j}$ being the j th ensemble member of $\mathbf{y}_i$. We define $\mathbf{y}_s > \mathbf{y}_t$ ($\mathbf{y}_s < \mathbf{y}_t$) if the probability that a randomly selected member of ensemble $\mathbf{y}_s$ exceeds a randomly selected member of ensemble $\mathbf{y}_t$ is larger (smaller) than 0.5. If the ensemble members of a forecast are interpreted as random samples from an underlying probability distribution, this definition is fully consistent with the conceptual decision rule proposed by MW09 for forecasts that are issued as continuous probability distributions (appendix A in MW09).

With this definition, two ensemble forecasts $\mathbf{y}_s$ and $\mathbf{y}_t$ can be ranked by the following algorithm:

- (i) Construct all possible pairs $\{y_{s,i}, y_{t,j}\}$, with $i \in \{1, \dots, m_s\}$ and $j \in \{1, \dots, m_t\}$.
- (ii) For each of these pairs determine the test statistic $q_{i,j}$ with $q_{i,j} = 1$ if $y_{s,i} > y_{t,j}$, $q_{i,j} = 0$ if $y_{s,i} < y_{t,j}$, and $q_{i,j} = 0.5$ if $y_{s,i} = y_{t,j}$.
- (iii) Calculate $F_{s,t} = \sum_i \sum_j q_{i,j} / m_s m_t$, which is the proportion of ensemble member pairs with $y_{s,i} > y_{t,j}$.
- (iv) Define: $\mathbf{y}_s > \mathbf{y}_t$ if $F_{s,t} > 0.5$, $\mathbf{y}_s = \mathbf{y}_t$ if $F_{s,t} > 0.5$, and $\mathbf{y}_s < \mathbf{y}_t$ if $F_{s,t} < 0.5$.

Note that $F_{s,t}$ and $F_{t,s}$ are statistically complementary (i.e., $F_{s,t} = 1 - F_{t,s}$). Also note that $\mathbf{y}_s = \mathbf{y}_t$ does not imply that $\mathbf{y}_s$ and $\mathbf{y}_t$ are *identical*, but rather that on the basis of these two forecasts it cannot be decided which of the two corresponding observations is likely to have the higher value. This can lead to situations that may appear paradoxical at first sight. Consider for example two hypothetical 3-member forecasts $\mathbf{y}_1 = (3, 3, 3)$ and $\mathbf{y}_2 = (2, 3, 10)$. The ranking algorithm defined above would yield $\mathbf{y}_1 = \mathbf{y}_2$, even though the forecasts are obviously not identical. In fact, intuitively one might argue

that $\mathbf{y}_2 > \mathbf{y}_1$ is more reasonable since the ensemble mean of $\mathbf{y}_2$ exceeds that of $\mathbf{y}_1$. However, without making additional assumptions concerning the underlying forecast distribution, there is no basis to rank $\mathbf{y}_1$ and $\mathbf{y}_2$. The fact that the distance between 3 and 2 (i.e., between the first members of $\mathbf{y}_1$ and $\mathbf{y}_2$) is smaller than the distance between 3 and 10 (i.e., between the third members of $\mathbf{y}_1$ and $\mathbf{y}_2$) becomes irrelevant since we do not know the statistical “closeness” of 2, 3, and 10; that is, we do not know the probability densities of the spaces between 2 and 3, and between 3 and 10. As a consequence of this, the logical operator “=” is not transitive (i.e., it is possible to find forecasts $\mathbf{y}_3$ such that $\mathbf{y}_1 = \mathbf{y}_2$ and $\mathbf{y}_1 = \mathbf{y}_3$, but $\mathbf{y}_2 \neq \mathbf{y}_3$). As an example, consider an additional forecast $\mathbf{y}_3 = (2, 3, 5)$ that satisfies $\mathbf{y}_3 = \mathbf{y}_1$ and $\mathbf{y}_3 < \mathbf{y}_2$. The lack of transitivity in this example is not a paradox. It simply reflects the fact that, while we do not know whether forecast value 2 is statistically closer to 3 than are forecast values 5 or 10 (i.e., $\mathbf{y}_1 = \mathbf{y}_2$ and $\mathbf{y}_1 = \mathbf{y}_3$), we do know that forecast value 10 exceeds forecast value 5, regardless of the underlying forecast densities (i.e., $\mathbf{y}_3 < \mathbf{y}_2$). Such lack of transitivity might be aesthetically disturbing, but it is irrelevant for the computation of D , since D is based on a serial assessment of forecast *pairs* only so that the order statistic outlined above is well defined. Having said that, in practice ensemble sizes of 20 and more are common, implying that ensemble forecasts are only rarely tied and violations of transitivity are unlikely to be observed frequently.

We now further simplify the formulation of the ranking procedure. Steps (i) to (iii) in the algorithm defined above are equivalent to a nonparametric test for the difference in central tendencies of two ensemble forecasts, namely the Mann–Whitney U test. By applying the equation for the Mann–Whitney U statistic (Sheshkin 2007), steps (i)–(iii) can be summarized in a single equation for $F_{s,t}$:

$$F_{s,t} = \frac{\sum_{i=1}^{m_s} r_{s,t,i} - \frac{m_s(m_s + 1)}{2}}{m_s m_t}, \quad (1)$$

with $r_{s,t,i}$ being the rank of $y_{s,i}$ with respect to the set of pooled ensemble members $\{y_{s,1}, y_{s,2}, \dots, y_{s,m_s}, y_{t,1}, y_{t,2}, \dots, y_{t,m_t}\}$, if sorted in ascending order. The second term in the numerator of Eq. (1) represents the sum of the ranks that would be obtained if all ensemble members of $\mathbf{y}_t$ exceeded those of $\mathbf{y}_s$, and so the numerator as a whole calculates the number of times that an ensemble member of $\mathbf{y}_s$ exceeds an ensemble member of $\mathbf{y}_t$. Thus, if all the ensemble members of $\mathbf{y}_t$ do exceed those of $\mathbf{y}_s$, the numerator will be 0, while if the converse is true the first

term in the numerator will equal $(m_s/2)(2m_t + m_s + 1)$, and so $F_{s,t}$ will be 1.

By repeated application of Eq. (1), it is now straightforward to determine R_s , the rank of forecast $\mathbf{y}_s$ within a set of n ensemble forecasts $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$:

$$R_s = 1 + \sum_{\substack{t=1 \\ t \neq s}}^n u_{s,t} \quad \text{with} \quad u_{s,t} = \begin{cases} 1 & \text{if } F_{s,t} > 0.5 \\ 0.5 & \text{if } F_{s,t} = 0.5 \\ 0 & \text{if } F_{s,t} < 0.5 \end{cases}. \quad (2)$$

We illustrate the application of Eqs. (1) and (2) with a simple example. Consider the three 5-member ensemble forecasts mentioned at the beginning of this section: $\mathbf{y}_1 = (22, 23, 26, 27, 32)$, $\mathbf{y}_2 = (28, 31, 33, 34, 36)$, and $\mathbf{y}_3 = (24, 25, 26, 27, 28)$. To determine their ranks R_1 , R_2 , and R_3 with Eq. (2), one needs to calculate $F_{1,2}$, $F_{2,1}$, $F_{1,3}$, $F_{3,1}$, $F_{2,3}$, and $F_{3,2}$. We exemplify the procedure for $F_{1,2}$. As a first step, the ensemble members of $\mathbf{y}_1$ and $\mathbf{y}_2$ are pooled together and sorted in ascending order, yielding $(22, 23, 26, 27, 28, 31, 32, 33, 34, 36)$. The ranks of the ensemble members of $\mathbf{y}_1$ with respect to this pooled vector are then determined: $r_{1,2,1} = 1$, $r_{1,2,2} = 2$, $r_{1,2,3} = 3$, $r_{1,2,4} = 4$, and $r_{1,2,5} = 7$. Using these values in Eq. (1) with $m_1 = m_2 = 5$ yields $F_{1,2} = 0.08$. Applying the same procedure with $\mathbf{y}_1$ and $\mathbf{y}_3$ ($\mathbf{y}_2$ and $\mathbf{y}_3$) yields $F_{1,3} = 0.44$ ($F_{2,3} = 0.98$). The corresponding transposes are $F_{2,1} = 0.92$, $F_{3,1} = 0.56$, and $F_{3,2} = 0.02$. Using these F values in Eq. (2) yields the following ensemble ranks: $R_1 = 1$, $R_2 = 3$, and $R_3 = 2$.

b. Formulations of D for ensemble forecasts

With this definition of how to rank ensemble forecasts, the ensemble version of D can be calculated in exactly the same way as if the forecasts were deterministic and measured on a continuous (or ordinal) scale (viz., by constructing all possible sets of two forecast–observation pairs and counting how often the observations can be correctly ranked by the forecasts; see Fig. 1). For this case, that is for forecasts that are deterministic and continuous, MW09 have derived formulations of D that depend on the ranks of the forecasts, but not on the actual forecast values [Eqs. (8), (18), and (22) in MW09]. Hence, once the ranks of the ensemble forecasts to be verified have been determined, these equations of MW09 can be equally applied to ensemble forecasts. Distinguishing between binary, categorical, and continuous observations, one obtains the following formulations for the ensemble version of D .

CASE 1: BINARY OBSERVATIONS [COUNTERPART OF EQ. (8) IN MW09]

$$D = \frac{\sum_{j=1}^{n_1} R_{1,j} - \frac{n_1(n_1 + 1)}{2}}{n_0 n_1}. \quad (3)$$

Here n_1 is the number of events and n_0 is the number of nonevents that have been observed. Here $R_{1,j}$ is the rank of that ensemble forecast that corresponds to the j th event that has been observed. The second term in the numerator represents the sum of the ranks for the worst possible set of forecasts for the events (i.e., the forecasts for the events are all ranked first), and so the numerator as a whole calculates how often a rank for the forecasts corresponding to an event is greater than for forecasts corresponding to a nonevent. We illustrate the meaning of $R_{1,j}$ with a simple example. Consider a set of 10 ensemble forecasts with ranks $\mathbf{R} = \{3, 1, 9, 7, 5, 4, 8, 2, 6, 10\}$, determined by Eq. (2), and corresponding binary observations $\mathbf{x} = \{0, 1, 1, 0, 0, 0, 1, 0, 1, 0\}$, with “1” indicating that an event has been observed. Consequently, one has $n_1 = 4$ and $n_0 = 6$. The 4 ensemble forecasts corresponding to an event have ranks of 1, 9, 8, and 6, implying that $R_{1,1} = 1$, $R_{1,2} = 9$, $R_{1,3} = 8$, and $R_{1,4} = 6$. Using these values in Eq. (3) yields $D = 0.58$.

CASE 2: CATEGORICAL OBSERVATIONS [COUNTERPART OF EQ. (18) IN MW09]

$$D = \frac{\sum_{k=1}^{c-1} \sum_{l=k+1}^c \left[\sum_{j=1}^{n_l} R_{l,k,j} - \frac{n_l(n_l + 1)}{2} \right]}{\sum_{k=1}^{c-1} \sum_{l=k+1}^c n_k n_l}. \quad (4)$$

Here c is the number of observed categories, and n_l denotes how often category $l \in \{1, \dots, c\}$ has been observed. Here $R_{l,k,j}$ has the following meaning: let the forecasts for when categories k or l have been observed be pooled and ranked in ascending order. Among this subset, $R_{l,k,j}$ denotes the rank of that ensemble forecast that corresponds to the j th observation in category l .

CASE 3: CONTINUOUS OBSERVATIONS [COUNTERPART OF EQ. (22) IN MW09]

$$D = \frac{1}{2}(\tau_{\mathbf{R}, \mathbf{x}} + 1). \quad (5)$$

Here $\tau_{\mathbf{R}, \mathbf{x}}$ is Kendall's rank correlation coefficient (Sheshkin 2007) between the n observations and the

n -element vector of corresponding ensemble ranks $\mathbf{R} = (R_1, \dots, R_n)$ as defined in Eq. (2).

3. Example

As an example, consider seasonal forecasts produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) Seasonal Prediction System 3 (Anderson et al. 2007). Hindcasts of mean near-surface (2 m) temperature, averaged over months December–February, have been used. Data stem from the ENSEMBLES project database (van der Linden and Mitchell 2009). All hindcasts have been started from 1 November initial conditions and cover the period 1960–2001. There are nine ensemble members. Verification is gridpointwise against data from the 40-yr ECMWF Re-Analysis (ERA-40) dataset (Uppala et al. 2005). The resulting skill maps are shown in Fig. 2. In Fig. 2a, the observations have been binned into two equiprobable categories (temperatures above and below average), and Eq. (3) for binary outcomes has been applied to calculate D . In Fig. 2b, the observations have been binned into three equiprobable categories, and Eq. (4) for categorical outcomes has been applied. Finally, in Fig. 2c, the “raw” observation values have been used, and D has been calculated with Eq. (5) for continuous observations.

In all three cases, the skill patterns obtained are consistent with earlier verification studies using different skill metrics (e.g., Weigel et al. 2008), showing that seasonal predictability of temperature is highest in the tropics, particularly the equatorial Pacific. Skill is seen to decrease systematically from binary to categorical to continuous observations. For instance, the skill average over the Niño-3.4 region (5°S – 5°N , 120° – 170°W) is $D = 0.97$ for binary observations, implying that in 97% of the cases the forecasts are able to correctly discriminate between the Niño-3.4 index being above and below average. If the observations are binned in three rather than two categories, skill drops to $D = 0.94$; and it is further reduced ($D = 0.87$) if continuous observations are considered. This loss of discriminative power can be explained by the additional precision that is required to discriminate between three rather than two categories, and even more so to discriminate between $n = 42$ discrete observations because then the forecasts have to successfully discriminate between some observations that differ only by small amounts.

4. Conclusions

This study has closed a gap in the verification framework of MW09 in providing formulations of the

FIG. 2. The generalized discrimination score D (%) for ECMWF System 3 forecasts of 2-m temperature as described in section 3. The observations have either been binned in (a) two equiprobable categories (“binary observations”), (b) three equiprobable categories (“categorical observations”), or (c) have not been binned at all (“continuous observations”).

generalized discrimination score D for ensemble forecasts. Discrimination is one of the most fundamental attributes of prediction skill in that it measures whether forecasts differ when their corresponding observations differ. While forecasts with high discriminative power may still be subject to systematic errors (e.g., bias, overconfidence) and may require (re)calibration to become useful (Weigel et al. 2009), forecasts lacking discrimination are useless by principle. Discrimination can therefore be considered as a necessary, but not sufficient attribute of prediction skill. It does not tell us how good a set of forecasts is if taken at face value, but rather how useful a set of forecasts can potentially be after appropriate calibration and postprocessing.

With the formulations of D provided here, it is possible to calculate discrimination for a set of ensemble forecasts without requiring that the ensemble members are transformed into probabilistic forecasts prior to verification. This has the advantage that the skill values obtained are not shadowed by potentially inappropriate assumptions concerning probabilistic ensemble interpretation. While some of the formulations presented here may appear “bulky” [e.g., Eq. (4)], their implementation is straightforward, and their interpretation follows the simple and intuitive principle introduced by MW09. As for all other formulations of D , the score is interpretable as an indication of how often the forecasts are correct, regardless of how many ensemble members there are, and regardless of whether the observed outcomes are measured on a binary, categorical or continuous scale. It has been argued that this property makes the score particularly useful for providing information of forecast quality to the general public.

Computer code, written in R, is available [<http://cran.r-project.org> (package “afc”)] for the procedures described here and in MW09. FORTRAN code is available from the authors upon request.

Acknowledgments. This study was funded by the Swiss National Science Foundation through the National Centre for Competence in Research (NCCR) Climate and by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA10OAR4310210). The views expressed herein are those of the authors and

do not necessarily reflect the views of NOAA or any of its subagencies.

REFERENCES

- Anderson, D., and Coauthors, 2007: Development of the ECMWF seasonal forecast system 3. ECMWF Tech. Memo. 503, 56 pp.
- Bröcker, J., and L. A. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus*, **60A**, 663–678.
- Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (roc) and levels (rol) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166.
- , and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349.
- Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.
- , and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasters*, **7**, 435–455.
- Pierce, C. S., 1884: The numerical measure of success in predictions. *Science*, **4**, 453–454.
- Sheshkin, D. J., 2007: *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 1776 pp.
- Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012.
- van der Linden, P., and J. F. B. Mitchell, Eds., 2009: ENSEMBLES: Climate change and its impacts at seasonal, decadal and centennial timescales—Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, 160 pp.
- Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134**, 241–260.
- , —, and —, 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Mon. Wea. Rev.*, **137**, 1460–1479.