The Probability Anomaly Correlation and Calibration of Probabilistic Forecasts

HUUG VAN DEN DOOL

NOAA/NWS/NCEP/Climate Prediction Center, College Park, Maryland

EMILY BECKER

NOAA/NWS/NCEP/Climate Prediction Center, College Park, and Innovim, LLC, Greenbelt, Maryland

LI-CHUAN CHEN

NOAA/NWS/NCEP/Climate Prediction Center, and Earth System Science Interdisciplinary Center/Cooperative Institute for Climate and Satellites, University of Maryland, College Park, College Park, Maryland

QIN ZHANG

NOAA/NWS/NCEP/Climate Prediction Center, College Park, Maryland

(Manuscript received 15 June 2016, in final form 3 November 2016)

ABSTRACT

An ordinary regression of predicted versus observed probabilities is presented as a direct and simple procedure for minimizing the Brier score (BS) and improving the attributes diagram. The main example applies to seasonal prediction of extratropical sea surface temperature by a global coupled numerical model. In connection with this calibration procedure, the probability anomaly correlation (PAC) is developed. This emphasizes the exact analogy of PAC and minimizing BS to the widely used anomaly correlation (AC) and minimizing mean squared error in physical units.

1. Introduction

This paper is about calibrating the probability forecast derived from an ensemble of numerical forecasts. The need for calibration arises because of systematic errors made by forecast models. There are at least two ways to assess the outcome of a forecast. The first approach, which is quite common, is to compare the ensemble of forecasts (usually as an ensemble mean) to the verifying datum in physical units. The second approach, explored here, is to compare predicted probabilities to the observed probability.[1] We design a regression method for the second approach. The paper is organized around the formal similarity of regression using variables in physical and probabilistic units, respectively.

One of the more prevalent skill metrics for tracking the skill of numerical weather prediction (NWP) is the anomaly correlation (AC). As a notion and as an accepted measure of skill, the AC has been around in the literature since about 1970 (Miyakoda et al. 1971). The word ''anomaly'' in the AC refers to both 1) the difference of the verifying observation/analysis and some reference state or climatology and 2) the difference of an actual forecast and that same climatology. Although used mainly as a metric to track skill over time or to compare model A to model B, a correlation always has a deeper interpretation that refers to the most basic verification attributes of all, the mean-square error (MSE): the correlation (a number in the range 0–1)[2] is proportional to

[1] Probability as such cannot be observed, so observed probability (usually 0 or 1, depending on whether an event occurs or not) is a bit of an abstraction.

[2] Strictly speaking, a correlation is in the range between −1 and +1, but in forecast verification on a sufficient sample significant negative correlation should not happen. Indeed, we do not ever contemplate changing the sign of a predicted anomaly.

Corresponding author e-mail: Huug van den Dool, huug.vandendool@noaa.gov

DOI: 10.1175/WAF-D-16-0115.1

the regression coefficient by which one should multiply the predicted anomalies by in order to minimize the MSE. When an NWP model is evaluated over many cases and is found to predict anomalies of the same magnitude as the observed anomalies, the predicted anomalies should be multiplied by the correlation itself, if minimizing MSE is the main goal[3]—in that circumstance the correlation equals the regression coefficient. When AC = 0, the forecasts should be damped all the way back to climatology in order to minimize MSE. Minimizing a cost function like MSE has been fundamental in many sciences for hundreds of years. The correspondence of the correlation and a skill score based on MSE as an attribute for accuracy was discussed at length in Murphy and Epstein (1989).

The point of this article is to extend the logic and purpose of the AC to the realm of probability forecasts. Probabilistic forecasting has started in earnest recently, when production of multiple runs to generate an ensemble of NWP forecasts became computationally feasible (Tracton and Kalnay 1993). As it turns out, there was already an MSE within the context of probability forecasts: it is called the Brier score (BS; Brier 1950), or, more generally, the probability score. We here present the probability anomaly correlation (PAC), which not only is a measure of skill in its own right as a single number (0–1), but also directly suggests that by damping probability anomalies the BS can be minimized. To our knowledge the PAC has not been exploited before. While the AC generally works with variables expressed in physical units, the PAC works with probabilities (i.e., fractions or percentages).

To simplify matters greatly, we characterize a probability forecast (a very detailed pdf in principle) in terms of probabilities for three terciles that are designed such that they climatologically happen 1/3 of the time. Nothing we present depends fundamentally on the number of classes used; three is traditional in some settings (Hamill et al. 2004). The probability anomaly would be the difference of the predicted probability from 1/3, which is the climatological probability for a tercile-based system.

Both in the case of AC and PAC we speak of damping (reducing anomalies via regression). In theory, the inflation of anomalies may be called for if (P)AC is high and the predicted anomalies are smaller than what would minimize the squared error term. However, this underconfidence, or the need for inflation in order to minimize the MSE or BS, rarely happens in practice.

A formal procedure to minimize a probability score appears to be rare. Dutton (2009; see his Fig. 9) appears to rotate lines in the attribute diagram closer to perfect reliability, but the formal method to do so is not specified. We are only aware of Gneiting et al. (2005), who minimize the continuous ranked probability score. Hamill et al. (2004) used logistic regression (using the ensemble mean as a predictor) to predict the probability of, for example, the above median tercile.

In section 2 we present the basic concept and equations, simple linear regression arguments, and the datasets used. A precise definition of what we mean by damping is given in section 2a. We then present an example in section 3 to show how well the PAC works. The example is for the prediction of sea surface temperature in the next month during 1982–2010 in the extratropical Northern Hemisphere (NH) by the Climate Forecast System version 2 (CFSv2) (Saha et al. 2014). The attractive reduction in BS is further illustrated by the usual attributes diagram (Hsu and Murphy 1986). Section 4 presents a few conclusions.

2. Methods and data

a. Equations

The traditional anomaly correlation is given by

$$\mathrm{AC} = \frac{\frac{1}{N}\sum_{i=1}^{N} F_i' O_i'}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} F_i'^2 \frac{1}{N}\sum_{i=1}^{N} O_i'^2}} \tag{1}$$

or

$$\frac{\mathrm{cov}(F, O)}{\mathrm{sd}_F \mathrm{sd}_O},$$

where F and O are the forecast and (verifying) observation in physical units (Wilks 2006; Van den Dool 2007). The prime represents a departure from a reference state C, usually observed climatology C_o; that is, $F' = F - C_o$ and $O' = O - C_o$. The summation is over time, $i = 1, \ldots, N$ with N pairs of forecasts and observations. As described, the AC is a regular correlation pointwise. Sometimes the AC also features a summation in space across grid points (not shown). When the climatology C used to form anomalies, via $F' = F - C$ and $O' = O - C$, is external to the sample at hand, the AC has a distinctly noncentered flavor (Van den Dool 2007).

[3] For statisticians it may seem obvious that MSE should be minimized, but practitioners in meteorology have a more practical concern: damping of forecast anomalies toward climatology may result in blank predicted weather maps at longer forecast times. The correspondence of formal metrics such as MSE and what forecasters want (on behalf of their customers) is always an issue. Here, the implicit properties of the (P)AC are actually helpful.

The resulting AC in that situation is still within normal bounds for a correlation, since all terms, the covariance cov(F, O) and the standard deviation of the forecast sd_F and observation sd_O, are evaluated with respect to the same C with appropriate summing over time, space, or both.

The mean square error is given by

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(F_i' - O_i')^2, \tag{2}$$

with the symbols as explained above. The primes are immaterial in (2) if the same C is subtracted from both F and O. Via standard regression, the MSE is minimized by creating an adjusted forecast as

$$F_i'^{\mathrm{adj}} = aF_i' + b, \tag{3}$$

where the regression coefficient a can be expressed as $a = \mathrm{cov}(F, O)/\mathrm{sd}_F^2$ or $\mathrm{AC} \times \mathrm{sd}_O/\mathrm{sd}_F$. The intercept b is zero when C is the sample mean. As stated in the introduction and expressed in (3), if $\mathrm{sd}_O = \mathrm{sd}_F$, one should mulitiply F_i' by AC in order to minimize MSE. Another option is to attempt this as a two-step process: the $\mathrm{sd}_O/\mathrm{sd}_F$ factor first corrects for the overall amplitude error in the forecasts F', followed by multiplication by AC if the lowest MSE is desired.

The minimized MSE can be calculated without explicitly having to amplitude calibrate forecast anomalies. The AC has this implicit quality and is telling us about skill (in the form of an MSE attribute) using the minimized MSE. The MSE of the climatological forecast (i.e., always predicting climatology) corresponds to the ''no skill'' situation.

We can now describe the PAC in exactly the same terms, using lowercase letters for probabilities instead of uppercase letters as above for physical units:

$$\mathrm{PAC} = \frac{\frac{1}{N}\sum_{i=1}^{N} p_i' o_i'}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} p_i'^2 \frac{1}{N}\sum_{i=1}^{N} o_i'^2}} \tag{4}$$

or

$$\frac{\mathrm{cov}(p, o)}{\mathrm{sd}_p \mathrm{sd}_o},$$

where the prime is the departure of p from the reference or climatological probability c_p of 1/3 (by design when using terciles); the index i goes across time from 1 to N; p is the predicted probability for a particular tercile; and o is 0 or 1 depending on the event (a hit of that tercile) happening or not. The BS is given by

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i' - o_i')^2. \tag{5}$$

One minimizes the BS by multiplying p_i' by a regression coefficient, which is proportional to PAC as follows:

$$p_i'^{\mathrm{adj}} = \mathrm{PAC}\frac{\mathrm{sd}_o}{\mathrm{sd}_p} p_i'. \tag{6}$$

We dropped the intercept in (6), assuming the sample mean of the primed quantities is zero.

The word ''damping'' is used throughout the paper. We define damping in general as multiplying $F' = F - C$ by a constant, usually in the [0, 1] range, therefore moving F closer to C in physical units. Similarly, we define damping within the probability context as multiplying $p' = p - c_p$ by a constant, usually in the [0, 1] range, moving the forecast p closer to c_p in terms of probability units. Equations (3) and (6) determine the amount of damping required.

In addition to the $\mathrm{BS}_{\mathrm{fct}}$ of the forecast method at hand calculated by (5), we require an expression for the Brier skill score (BSS) as follows:

$$\mathrm{BSS} = (\mathrm{BS}_{\mathrm{control}} - \mathrm{BS}_{\mathrm{fct}})/\mathrm{BS}_{\mathrm{control}}, \tag{7}$$

where $\mathrm{BS}_{\mathrm{control}}$ is the Brier score of always predicting climatology, which, in this case means always predicting 1/3.

The BS can be decomposed in several terms, including resolution (RES), reliability (REL), and an ''uncertainty'' term U (Wilks 2006). Symbolically, BS = REL − RES + U. The REL and RES terms are illustrated in an attributes diagram (Hsu and Murphy 1986). In forecast verification, reliability represents the comparison of a forecast probability for an event to the observed frequency of that event. For example, for a reliable forecast, all forecasts of 40% probability of above normal T2m should be observed 40% of the time. Resolution indicates the use of different forecast probabilities: the ability of the forecast system to assign probabilities different from the climatological probability. The third term U depends only on the observations and is not discussed further.

One can entertain various extensions, like summing over all terciles, summing in space, or using fractional numbers for o_i, instead of 0 or 1 (Candille and Talagrand 2008; Chen et al. 2017).

b. *Data: Model forecasts*

The forecasts are extracted for the period 1982–2010 from the dataset produced by the CFSv2 model (Saha et al. 2014). We are studying lead +1-monthly retroactive forecasts available at 1° latitude × 1° longitude resolution for sea surface temperature (SST). All

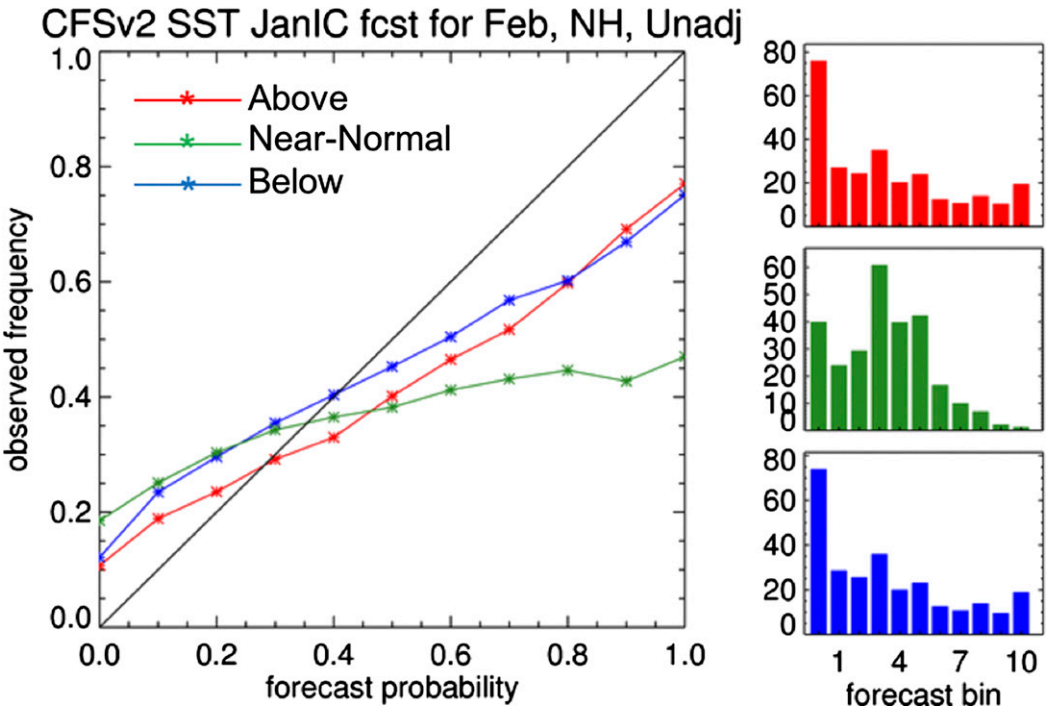

FIG. 1. Attributes diagram of CFSv2 forecasts of monthly mean SST during February 1982–2010 starting from initial conditions in early January. Results are aggregated for all grid points over the extratropical oceans between 25° and 75°N. The probabilities are divided into 11 bins: <5%, 5%–15%, 15%–25%, . . . , 85%–95%, and >95%. Each cross in the left panel gives the observed frequency as a function of the predicted probability. Blue, green, and red lines are for three terciles representing below normal, near normal, and above normal. The three histograms on the right report the total number of forecasts in each bin in thousands. Example: there were about 75 000 occasions with a 0%–5% probability for the below-normal tercile [i.e., only 0 or 1 ensemble members (out of 24) in that tercile].

datasets are global but are used here for certain domains specified below. Lead +1 means that, for example, forecasts for the month of February are generated from initial conditions near 1 January.

c. *Data: Verification fields*

For SST verification we use the Optimum Interpolation Sea Surface Temperature, version 2, analysis of Reynolds et al. (2002). This analysis uses both satellite data and in situ records from ships and buoys. The native resolution of the Reynolds et al. (2002) SST is 1° latitude × 1° longitude. The observation period used in this study runs from 1982 to 2010.

3. Example

In the example we consider the prediction of extratropical SST by the CFSv2 during 1982–2010 (29 yr). The target is the February monthly mean SST at all grid points over the ocean between 25° and 75°N. For each year there are 24 total initial states, from four sequential runs (initialized at 0000, 0600, 1200, and 1800 UTC) on each of six days, once every pentad in late December and early January. We choose this SST example as being representative of monthly or seasonal prediction of varying lead times in general; extratropical SST is not predicted as well as tropical SST in the Pacific, but the results are much better than those for air temperature and, especially, for precipitation over land.

Forecast probabilities are formed from the retrospective forecasts in the following manner. For each of the 29 yr in the sample, the thresholds delimiting the upper and lower thirds of the historical record (terciles) are calculated from the other 28 yr. The sample size is 628; that is, 24 ensemble members times 28 yr. For example, the tercile thresholds for the 1982 test case are calculated from the 1983–2010 hindcasts. Then, the 24 ensemble members from the test year are sorted into above-normal, near-normal, and below-normal categories according to the tercile thresholds. The percentage of ensemble members falling in each tercile is the forecast probability; this is the "count" method.

Figure 1 aggregates across the entire grid the verification results of the raw or unadjusted probabilities for

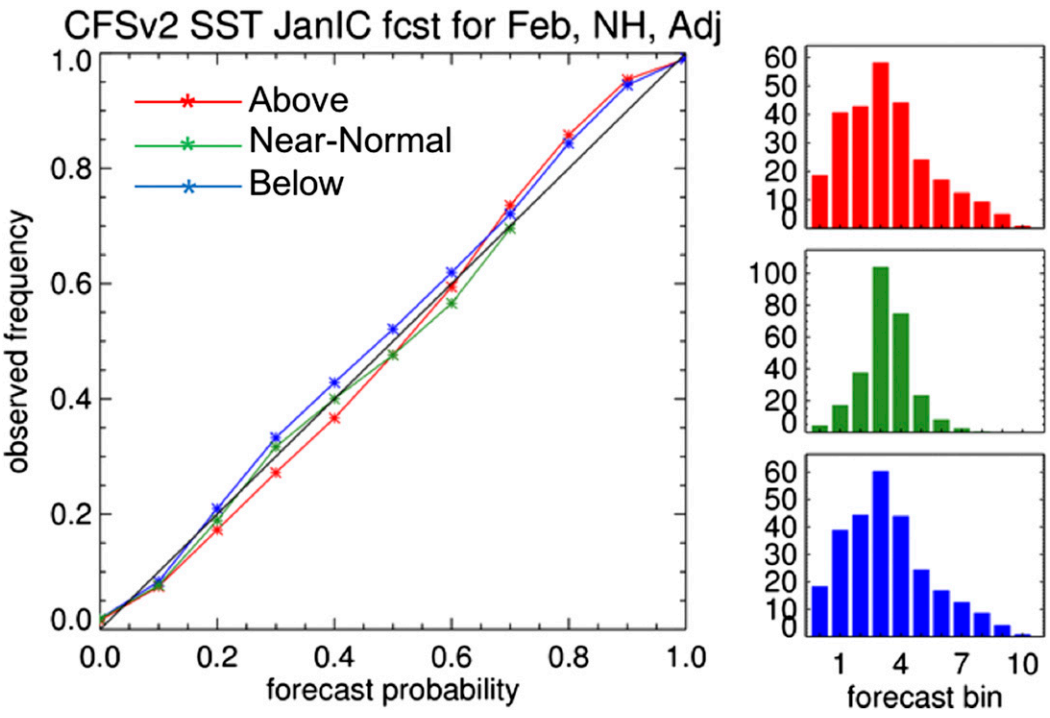

FIG. 2. As in Fig. 1, but after PAC adjustment.

the three terciles separately; the colors red, blue, and green refer to forecasts for above, below, and normal terciles, respectively, and are sometimes abbreviated as A, B, and N. The results here are familiar (e.g., Barnston et al. 2003; Palmer et al. 2005, 2008). First, the forecasts are overconfident, which means, for example, that an event conditioned on a predicted probability of 60% is verified to happen only about 40% of the time. All three lines have a slope of less than 45° (the line indicating a perfect correspondence between forecast probabilities and event occurrence—a ''reliable'' forecast). Second, for lack of skill, the overconfidence problem is the most severe for the middle tercile. This too is well known (Van den Dool and Toth 1991; Kharin et al. 2009). The inset histograms on the right show how often the 11 probability bins (bin 0 counts all forecasts of 0%–5%, bin 1 all forecasts of 5%–15%, etc.) are used: for each tercile (A, B, and N) each bin shows how many forecasts in that range were issued, aggregated for all years, for all grid points. It is possible that the CFSv2 system is too bullish (we see U-shaped histograms for the outer terciles, indicating a disproportionate number of forecasts are for the very highest or lowest probabilities) and assigns at times unrealistically high/low probabilities that need to be tempered in view of the modest overall skill.

One note of caution: although we call the situation depicted in Fig. 1 raw or unadjusted, we did in fact correct already for systematic errors in the mean and distributional aspects of the models, since the tercile thresholds are derived from the model's hindcast data before applying the count method. The verification terciles, which are derived from observations, may be very different from the model's tercile boundaries. If we had assigned model forecasts to observed terciles, Fig. 1 would look substantially worse, because the probabilistic verification scores would be negatively impacted not only by the mean bias, but also by the wrong standard deviation or higher-order pdf problems. The PAC adjustment and its benefits that we present in this study take Fig. 1 as the starting point. In Fig. 1 the model (CFSv2) looks flawed, but the results are not terrible.

Figure 2 is the same as Fig. 1, but after the PAC adjustment, where (4)–(6) have been applied. As we will see below, the BS was indeed reduced (minimized by linear means). We did not have a clear a priori sense of how the reduction in BS would work out in the attributes diagram, but Fig. 2 confirms a favorable result. For all three classes, the lines in the forecast probability versus observed frequency plot are close to the 45° line. This means ''reliability'' (one of the terms in the BS decomposition) is near perfect and makes a near-zero contribution to BS. The resolution has also improved. The latter can be judged visually from the angle of these three lines with the horizontal, which is higher in Fig. 2 than in Fig. 1; the higher the angle, the higher the resolution. The minimum BS is thus at 45°, where the

resolution is as great as possible without harming the reliability. [We do not offer "proof" that this is so, but see the comments by Mason (2012) concerning exchanging reliability for resolution or vice versa.] A beneficial impact of the PAC approach is also seen in the histograms on the r.h.s. of Fig. 2. The frequency of occurrence of the extreme bins (<5% and >95%) has been reduced, and the U-shape has turned into a bell shape, so forecasts for an extreme probability are fewer in number (compared with Fig. 1), but not zero. Since skill is low for the N class, its histogram has been pulled inward into the 25%–35% bin (bin 3 in the histogram figures) more so than for the A and B classes. But even in the N class, we are allowed an occasional probability anomaly of 5% or 10%. This is only allowed when the PAC > 0 (i.e., the method surmises inherent skill), even when BS skill for the raw forecast is negative. When PAC is zero, or negative (this has to be due to sampling) the PAC-adjusted prediction is set to the climatological probability of 1/3, a frequent occurrence for the N tercile, as shown by the high number of forecasts in this bin (Fig. 2, center histogram).

The area-aggregated PAC [i.e., summing (4) in space; not shown] for this example is about 0.4 for the A and B terciles, and 0.15 for the N class, but the PAC varies greatly in space. Values of 0.4 might give the impression of strong damping being required. However, the standard deviation from the observations is often larger than that of the forecast (e.g., Becker et al. 2014). Since sd_o is usually on the order of 50% larger than sd_f, and the actual damping factor is relative to this ratio [(6)], the damping is not as large as PAC alone indicates. The high sd_o/sd_f ratio for probabilities is largely caused by observations being either 0 or 1, or probability anomalies being either 1/3 or 2/3, while the forecasts are any fractional number, and their probability anomalies are smaller. So a PAC of 0.4 indicates a damping of the probability anomaly by about 0.6.

Table 1 shows the BS and BSS (Brier skill score) for this example. It is evident that the BS is reduced, and by about the same amount in all three terciles, when the PAC adjustment is executed. The improvements expressed as per BSS, that is, the improvement over a purely climatological forecast, look quite good for the A and B classes. The more noticeable improvement, however, is for the N class. For the raw forecasts the BSS is negative, because the BS is worse than the climatological control. Applying the PAC adjustment turns out to be more beneficial than just allowing the avoidance of embarrassment, which would be to bring the negative BSS up to zero. The PAC is positive for the N class, which means that with sufficient damping the BSS can actually be turned positive, and the histograms on the right in Fig. 2 do not have to be delta functions right at 0.333.

A nuisance problem with PAC-corrected probabilities is that the regression, as formulated, does not force probabilities for any tercile to be ≥0 or ≤100, nor is the sum of the probabilities of the three terciles forced to be 100%. Most of the time, these violations are inconsequential, but sometimes discrepancies on the order of 5%–10% occur. The appendix describes an iterative procedure for correcting these problems. The iteration always converges. This is called adjusted+ in the bottom row of Table 1. In no case does the BS suffer from pushing probabilities inside bounds; in fact, the BS is lowered by a tiny amount in all cases. Although BS was already minimized, it was minimized by linear means, and this does not preclude it from being lowered further. We did not pursue a more formal approach that would guarantee a priori proper behavior for PAC-adjusted probabilities; see Glahn (2014) for comments on such problems.

TABLE 1. Tabulations of BSs and BSSs by tercile for the 1-month lead prediction of monthly mean SST prediction aggregated for all grid points in NH extratropical oceans. Raw means forecasts from CFSv2 are assigned to terciles by the count method. Adjusted means that probability anomalies are regressed toward zero by the PAC method. For the definition of adjusted+, see the text.

	BS × 100			BSS × 100		
	A	N	B	A	N	B
Raw	18.7	23.5	20.1	14.6	−6.8	11.5
Adjusted	16.4	20.6	17.7	25.1	6.5	22.1
Adjusted+	16.3	20.4	17.4	25.7	7.4	23.2

The SST forecast above simply serves as an illustration; other variables show the exact same behavior. A more complete study (all variables, all start months) is forthcoming, with the baseline (raw or unadjusted) already described in Becker and Van den Dool (2016).

4. Conclusions and discussion

A method is proposed to calibrate raw forecast probabilities coming from a dynamical model or other prediction methods. It is simply a regression between predicted and observed probabilities, which should by definition minimize the Brier score. The name probability anomaly correlation (PAC) was chosen to stress the analogy of the traditional anomaly correlation and its relation to minimizing MSE. The PAC method is simple, direct, and appears to be successful on the examples studied. It is, however, unusual to think of regression applied directly to probabilities.

There are at least two reasons why probabilities coming straight from a model need some damping and/or smoothing. The first reason is noise. The most typical situation is one of a model being overconfident, especially when too few ensemble members are generated. The situation is aggravated by the count method, which has large round-off errors. In the limit of a single

ensemble member, the single outcome will be mapped (in the count method) as two 0s and one 1 for the predicted probability of the three classes. This would be a silly forecast. Unless inherent skill is perfect, the 0s and 1 should be moved toward ⅓; that is, the probability anomalies (which were either $-⅓$ or $+⅔$ in the example of a single ensemble member) are to be damped toward zero to improve the BS. The PAC tells us exactly by how much. This damping is close to smoothing that can be achieved in many other ways. Any procedure that reduces a 1 (''certainty'') to $1 - \varepsilon$, and giving the remaining ε to a neighboring class (or classes), is likely to make probability forecasts look better. It follows logically that, everything else being the same, a model with small ensemble size has more to gain from PAC-based probability adjustments than a model with a large ensemble. This consideration is especially important when comparing a single model to a multimodel ensemble.

A second reason, which is more subtle and profound, for damping is that counting model outcomes give an impression of potential predictability under the perfect model assumption: the degree to which a model is able to predict itself, thus illustrating the divergence of forecasts over time and the potential growth of errors (Lorenz 1982). As shown in Becker et al. (2014), predictability (where the ensemble of $N - 1$ members is used to predict one withheld member) tends to be higher than prediction skill (N-member ensemble mean versus the observation) for seasonal prediction. The PAC measures prediction skill—information that can be used to appropriately reduce predictability-based probabilities.

Although we claim that the PAC approach of calibrating probabilities is novel, we do not claim that this method is necessarily the best. Being ''best'' depends on the criteria and metrics used, and there must be at least a half-dozen postprocessing methods that clean up raw probabilities satisfactorily. Among them are the aforementioned Gneiting et al. (2005) approach, ''ensemble regression'' (Unger et al. 2009), and a regression method described in Tippett et al. (2014). The latter two methods design a regression between the ensemble mean and the observations, minimizing MSE in physical units at the start. Neither optimizes a probabilistic score by design. Compared with methods that only use the ensemble mean as predictor (Hamill et al. 2004; Tippett et al. 2014) the PAC method distinguishes itself by using predicted probabilities as predictors, in theory using distributional information. Whether this helps remains to be demonstrated.

As of April 2016 the PAC method was implemented and applied in real time to the North American Multimodel Ensemble (NMME); examples can be seen online (http://www.cpc.ncep.noaa.gov/products/NMME/). While the present article included an example for a single dynamical model, the NMME includes seven or eight models. There are various ways of going from a single to multiple PAC-adjusted models. This will be described in a forthcoming paper. Here, we present the idea of the PAC adjustment, and, as an example, we have applied it to a single model.

Acknowledgments. The manuscript was read by Malaquias Pena, Michelle L'Heureux, Stephen Baxter, John A. Dutton, Anthony Barnston, Michael Tippett, Richard P. James, and three anonymous reviewers. Their comments are gratefully acknowledged. This work was completed at the Climate Prediction Center. Support was given by NOAA's MAPP Predictions Program via Grant NA14OAR4310188.

APPENDIX

Final Alterations to Adjusted Probabilities

Alterations may need to be made to address three types of violations of probabilities listed below. If necessary, steps 1–3 are executed more than once.

1) When the adjusted probability p_k for class k is negative, we define a discrepancy (a negative number) as $p_k - 0.01$. Half the discrepancy is added to the two other classes, and p_k is set equal to 0.01. (Example: if the PAC-adjusted $p_1 = -0.05$, then p_1 is set to $+0.01$, and -0.03 is added to both p_2 and p_3.) This rule is applied to all classes in turn, $k = 1, 2, 3$, when applicable.
2) When the adjusted probability p_k for class k is larger than 100%, we define a discrepancy (a positive number) as $p_k - 0.99$. Half the discrepancy is added to the two other classes, and p_k is set equal to 0.99. This rule is applied to all classes $k = 1, 2, 3$, when applicable.
3) Third, we define a discrepancy as $(p_1 + p_2 + p_3) - 1$. One-third of this discrepancy (positive or negative) is subtracted from p_k, $k = 1, 2, 3$.

We keep repeating these three steps until no more action is needed. Two iterations generally appear to be enough.

REFERENCES

Barnston, A. G., S. J. Mason, L. Goddard, D. G. Dewitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.*, **84**, 1783–1796, doi:10.1175/BAMS-84-12-1783.

Becker, E., and H. Van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, doi:10.1175/JCLI-D-14-00862.1.

——, ——, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, doi:10.1175/JCLI-D-13-00597.1.

Brier, G. W., 1950: Verification of forecasts expressed in probabilities. *Bull. Amer. Meteor. Soc.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 959–971, doi:10.1002/qj.268.

Chen, L.-C., H. Van den Dool, E. Becker, and Q. Zhang, 2017: ENSO precipitation and temperature forecasts in the North American Multimodel Ensemble: Composite analysis and validation. *J. Climate*, doi:10.1175/JCLI-D-15-0903.1, in press.

Dutton, J. A., 2009: Weather, climate, and the energy industry. *Management of Weather and Climate Risk in the Energy Industry*, A. Troccoli, Ed., NATO Science for Peace and Security Series C: Environmental Security, Springer, 3–23.

Glahn, B., 2014: A nonsymmetric logit model and grouped predictand category development. *Mon. Wea. Rev.*, **142**, 2991–3002, doi:10.1175/MWR-D-13-00300.1.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, doi:10.1016/0169-2070(86)90048-8.

Kharin, V. V., Q. Teng, F. W. Zwiers, G. J. Boer, J. Derome, and J. S. Fontecilla, 2009: Skill assessment of seasonal hindcasts from the Canadian Historical Forecast Project. *Atmos.–Ocean*, **47**, 204–223, doi:10.3137/AO1101.2009.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34A**, 505–513, doi:10.1111/j.2153-3490.1982.tb01839.x.

Mason, S., 2012: Do statistical models trade resolution for reliability? *Seminar on Seasonal Prediction: Science and Applications*, Reading, United Kingdom, ECMWF, 73–82.

Miyakoda, K., R. F. Strickler, C. J. Nappo, P. L. Baker, and G. D. Hembree, 1971: The effect of horizontal grid resolution in an atmospheric circulation model. *J. Atmos. Sci.*, **28**, 481–499, doi:10.1175/1520-0469(1971)028<0481:TEOHGR>2.0.CO;2.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–582, doi:10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193, doi:10.1146/annurev.earth.33.092203.122552.

——, F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell, 2008: Toward seamless prediction. *Bull. Amer. Meteor. Soc.*, **89**, 459–470, doi:10.1175/BAMS-89-4-459.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Tippett, M. K., T. DelSole, and A. G. Barnston, 2014: Reliability of regression-corrected climate forecasts. *J. Climate*, **27**, 3393–3404, doi:10.1175/JCLI-D-13-00565.1.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, doi:10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Unger, D., H. Van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379, doi:10.1175/2008MWR2605.1.

Van den Dool, H. M., 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 215 pp.

——, and Z. Toth, 1991: Why do forecasts for near normal often fail? *Wea. Forecasting*, **6**, 76–85, doi:10.1175/1520-0434(1991)006<0076:WDFFNO>2.0.CO;2.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.