Adv. Geosci., 16, 137–142, 2008
www.adv-geosci.net/16/137/2008/

Advances in Geosciences

Bias Adjusted Precipitation Threat Scores

F. Mesinger

NCEP/Environmental Modeling Center, Camp Springs, Maryland, and Earth System Science Interdisciplinary Center (ESSIC) University of Maryland, College Park, Maryland, USA

Received: 16 December 2007 – Revised: 30 January 2008 – Accepted: 9 February 2008 – Published: 9 April 2008

Abstract. Among the wide variety of performance measures available for the assessment of skill of deterministic precipitation forecasts, the equitable threat score (ETS) might well be the one used most frequently. It is typically used in conjunction with the bias score. However, apart from its mathematical definition the meaning of the ETS is not clear. It has been pointed out (Mason, 1989; Hamill, 1999) that forecasts with a larger bias tend to have a higher ETS. Even so, the present author has not seen this having been accounted for in any of numerous papers that in recent years have used the ETS along with bias "as a measure of forecast accuracy".

A method to adjust the threat score (TS) or the ETS so as to arrive at their values that correspond to unit bias in order to show the model's or forecaster's accuracy in *placing* precipitation has been proposed earlier by the present author (Mesinger and Brill, the so-called dH/dF method). A serious deficiency however has since been noted with the dH/dF method in that the hypothetical function that it arrives at to interpolate or extrapolate the observed value of hits to unit bias can have values of hits greater than forecast when the forecast area tends to zero. Another method is proposed here based on the assumption that the increase in hits per unit increase in false alarms is proportional to the yet unhit area. This new method removes the deficiency of the dH/dF method. Examples of its performance for 12 months of forecasts by three NCEP operational models are given.

1 Introduction

Threat score (TS), also known as the critical success index (CSI, e.g., Schaefer, 1990); or equitable threat score (ETS) which is a modification of the threat score to account for the correct forecasts due to chance (Gilbert, 1884), is used at the National Centers for Environmental Prediction (NCEP) almost exclusively as the primary variable for verification of the skill in precipitation forecasting. It is almost always used along with the bias score. The use of these two measures is quite widespread also outside NCEP. The meaning of the ETS beyond its mathematical definition is however not all that clear. Yet, an understanding that the higher the ETS the better the model skill is for the particular threshold seems to prevail, and is spelled out precisely in this way in at least one case.

Several authors however have pointed out that a higher TS or ETS is not necessarily synonymous with more accurate forecasts. Thus, Shuman (1980) says "It is well known ... that ... higher threat scores can be achieved by increasing the bias above unity". Citing Mason (1989), Hamill (1999) writes "typically, the forecast with the larger bias (the wetter forecast) tends to have a higher ETS than if the two models had the same bias".

As a remedy, Shuman (1980) proposed a modified TS arrived at by assuming that both the forecast and the observed precipitation each cover a single circular area displaced relative to each other. Knowing the forecast area, F, correctly forecast area ("hits"), H, and the observed area, O, suffices to calculate the displacement. Keeping the displacement constant Shuman next assumed a reduction of the larger circle area to that of the smaller. Modified threat score can then be calculated. Shuman wanted a method that is using only a single set of values of the F, H, and O, and states that his aim is "to remove the effect of bias in overforecasting, and likewise to exact a penalty for underforecasting."

Hamill (1999) pointing out that comparisons of common threat scores like the ETS "are suspect unless the biases of competing forecasts are similar" adjusted verification thresholds of one model so as to equal biases of the other, by essentially relabeling the forecast contours of the former. Adjusting or relabeling forecast values of both or of more models to achieve biases of unity is an obvious possibility also.

The idea of the dH/dF method (Mesinger and Brill, 2004) was to use an interpolation or extrapolation function $H(F)$ hypothesizing how would the hit area change if the model bias at the considered threshold were changing, and use this function to obtain the value of H that accordingly the model would have if it had no bias. Since an ETS obtained using

Correspondence to: F. Mesinger
(fedor.mesinger@noaa.gov)

Published by Copernicus Publications on behalf of the European Geosciences Union.

this adjusted value of H presumably would have no influence of model bias by default the only influence remaining would be one of the *placement* of forecast precipitation.

There is of course a whole class of the so-called entity- and/or object-based methods (e.g., Ebert and McBride, 2000; Davis et al., 2006; many more) that strive to identify precipitation "events" and result in information that includes measures of the placement accuracy of the forecasts of these events. In contrast, the objective of the dH/dF method was to arrive at a placement accuracy measure without involving decisions needed for the identification of individual precipitation events. The adjusted ETS was intended to serve as an improvement on the popular ETS score, by not losing much on its simplicity of use while at the same time rendering the obviously important information on the placement accuracy of the forecasts, not clearly visible from a combination of the standard ETS and bias scores.

2 Method

Using the F, H, O notation, the threat score, TS, and the equitable threat score, ETS, are defined by

$$\mathrm{TS} = \frac{H}{F + O - H} \tag{1}$$

and

$$\mathrm{ETS} = \frac{H - FO/N}{F + O - H - FO/N} \tag{2}$$

respectively, with N denoting the total number of verification points, or events. Given a set of known values of F, H, and O, the interpolation or extrapolation function $H(F)$ of the dH/dF method was arrived at by postulating

$$\frac{dH}{dF} = a(O - H), \quad a{=}\text{const} \tag{3}$$

and requiring that the resulting function $H(F)$ satisfy the requirements:

number of hits H must be zero for F=0;

the function $H(F)$ has to satisfy the known value of H for the model's F, and,

$H(F)$ should remain less than but has to approach O as F increases.

These requirements led to the bias adjusted value of H

$$H_a = O\left(1 - \left(\frac{O - H}{O}\right)^{\frac{O}{F}}\right) \tag{4}$$

Using this value instead of the actual H, and O in place of F in Eq. (1) or (2), the bias adjusted threat or equitable threat score was calculated (Mesinger and Brill, 2004; see also Baldwin and Kain, 2006).

The performance of the dH/dF method when applied to TS in comparison with that of five other measures has been analyzed by Baldwin and Kain (2006) for an idealized case of circular forecast and observed precipitation areas, with results generally favorable to the dH/dF method. It was noticed however later that the function $H(F)$ resulting from the dH/dF scheme can have values $H > F$ near the origin, which is physically unreasonable. Needless to say this also makes the use of the function suspect when the problem does not actually happen.

To prevent this possibility an additional requirement to have $H(F) < F$ for all F is needed. This will be achieved if Eq. (3) is replaced by

$$\frac{dH}{dA} = b(O - H), \quad b = \text{const} \tag{5}$$

where $dA{=}dF{-}dH$ stands for additional false alarms occurring as F is increased by dF. It should be stressed that no claims are made to have b a physical constant of the forecasting system in place; the assumption is made merely to arrive at a single use interpolation/extrapolation function satisfying the requirements made.

While not as straightforward as that of Eq. (3), the solution of Eq. (5) is readily obtained via symbolic mathematical software, such as Mathematica or Matlab. One finds

$$H(F) = O - \frac{1}{b}\mathrm{lambertw}\left(bO\, e^{b(O-F)}\right) \tag{6}$$

where lambertw is the inverse function of

$$z = we^{w} \tag{7}$$

In Mathematica, it is denoted ProductLog; the name omega function is also used. Thus,

$$\mathrm{lambertw}(z) = w \tag{8}$$

Requiring that Eq. (6) satisfy the known values of F, H denoted as F_b, H_b, and using Eq. (7) and Eq. (8), one obtains

$$b = \frac{1}{F_b - H_b}\ln\left(\frac{O}{O - H_b}\right) \tag{9}$$

Thus,

$$H(F) = O - \frac{F_b - H_b}{\ln\left(\frac{O}{O-H_b}\right)}\mathrm{lambertw}$$
$$\left(\frac{O}{F_b - H_b}\ln\left(\frac{O}{O - H_b}\right)\left(\frac{O}{O - H_b}\right)^{\frac{O-F}{F_b-H_b}}\right) \tag{10}$$

is the required function $H(F)$ of the present method, to be referred to as dH/dA. Using it to obtain the bias adjusted value of H, H_a, we have

$$H_a = O - \frac{F - H}{\ln\left(\frac{O}{O-H}\right)}\mathrm{lambertw}\left(\frac{O}{F - H}\ln\left(\frac{O}{O - H}\right)\right) \tag{11}$$

where the subscripts of F_b and H_b have been omitted. The dH/dA bias adjusted values of the threat and the equitable

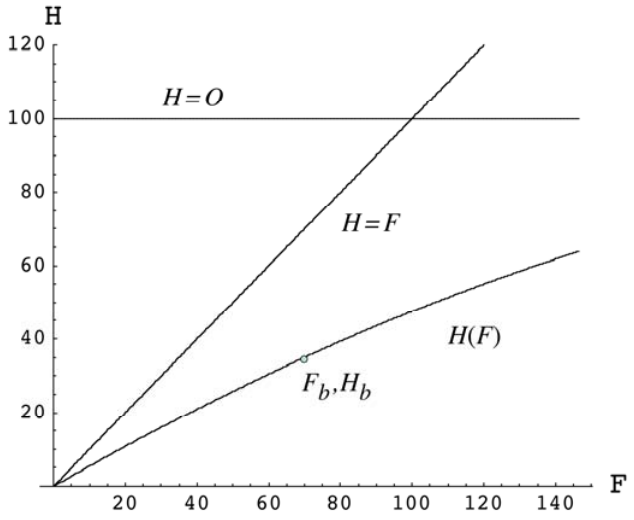

Fig. 1. Illustration of the bias adjustment method. A triplet of F, H, O values is considered known; since F, H are values for a forecast presumably having a bias, they are denoted by F_b, H_b. Values of F, H, O equal to 70, 35, and 100, respectively, are used for the plot shown. A function is sought $H(F)$ that will be consistent with the known asymptotics of the problem, and will enable extrapolation or interpolation of H to the value of F/O=1.

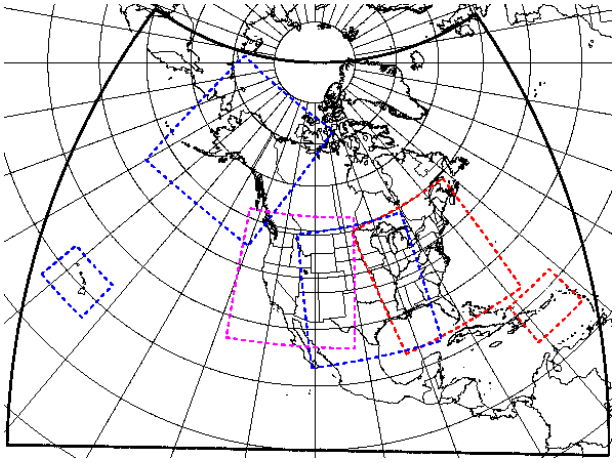

Fig. 2. Domains of the Eta 12 km operational model during the time of the scores shown in Figs. 3–5, February 2004–January 2005, heavy black line, and of the "high resolution windows" of the nested NMM model (Nonhydrostatic Mesoscale Model), dashed color lines. (Plot courtesy of Eric Rogers).

threat score, respectively, are now obtained using Eq. (11) as opposed to Eq. (4) for H while replacing F by O in Eq. (1) and Eq. (2).

An example of the resulting function $H(F)$ is shown in Fig. 1. It shows the shape of $H(F)$ for values of 70, 35, and 100, for the F_b, H_b and O, respectively. The bias adjusted value of H obtained is 47.5579. Standard ETS in this case is 0.2586, while the bias adjusted ETS is equal to 0.3112.

While lambertw or ProductLog function is not available with standard programming languages, ready-made codes are freely downloadable via the Internet that can be used for that purpose. For example, two codes are available at http://www.netlib.org/toms/443. The one named "WEW_B" was used for the real forecast examples of the next section.

A problem one can have with the method is that of the existence of singular cases. Three possible singular cases can be identified: (a) no false alarms (H_b=F_b), (b) all of the observed points have been hit (H_b=O), and (c) there were no hits (H_b=0). It is suggested that it is in the spirit of the method to declare in the former two cases H_a=O, given that in these two cases, (a) and (b), there is no basis to object to position accuracy. In the third case, that of hits H_b=0, obviously H_a should be set to zero as well.

3 Examples: 12 months of three NCEP models' scores

Precipitation scores of three NCEP models operational during the 12-month period February 2004–January 2005 are used to depict the impact of the method. One is the Global Forecast System (GFS), containing the NCEP operational global model, referred to now as GFS (e.g., GCWM Branch, EMC, 2003). The other is the Eta model (e.g., Pielke, 2002, pp. 542–544), run at the time at 12 km horizontal resolution. The Eta was obtaining its lateral boundary conditions from the GFS run of 6 h earlier. As of summer 2002 and through the period addressed, a still higher resolution model, NMM (Nonhydrostatic Mesoscale Model, e.g., Janjić, 2003) was run over six "high resolution windows" covering the contiguous United States (CONUS) area, Alaska, Hawaii, and Puerto Rico. The domains of the Eta and the NMM models are shown in Fig. 2. Over its three CONUS domains, the NMM was run at 8 km horizontal resolution. The Eta and the NMM both used 60 layers in the vertical and they have used essentially the same precipitation schemes.

Two of the CONUS domains of Fig. 2 were chosen for verification examples; the one centered over the eastern US (to be referred to as "East"), and that centered over the western US (to be referred to as "West"). The motivation for choosing the two domains is that over the East the impact of topography is less localized than in the mountainous West where it can dominate precipitation placement so that differences in the relative performance of models over the two regions may provide insights of interest. The period specified is on the other hand chosen as particularly attractive as it includes three months of by far the heaviest rains in the West during all of the period January 2004–August 2005, when the scores of these three models ceased being simultaneously available. Several high impact very heavy rain events in the West are included in the period selected, such as that having led to the La Conchita, CA, mudslide of January 2005.

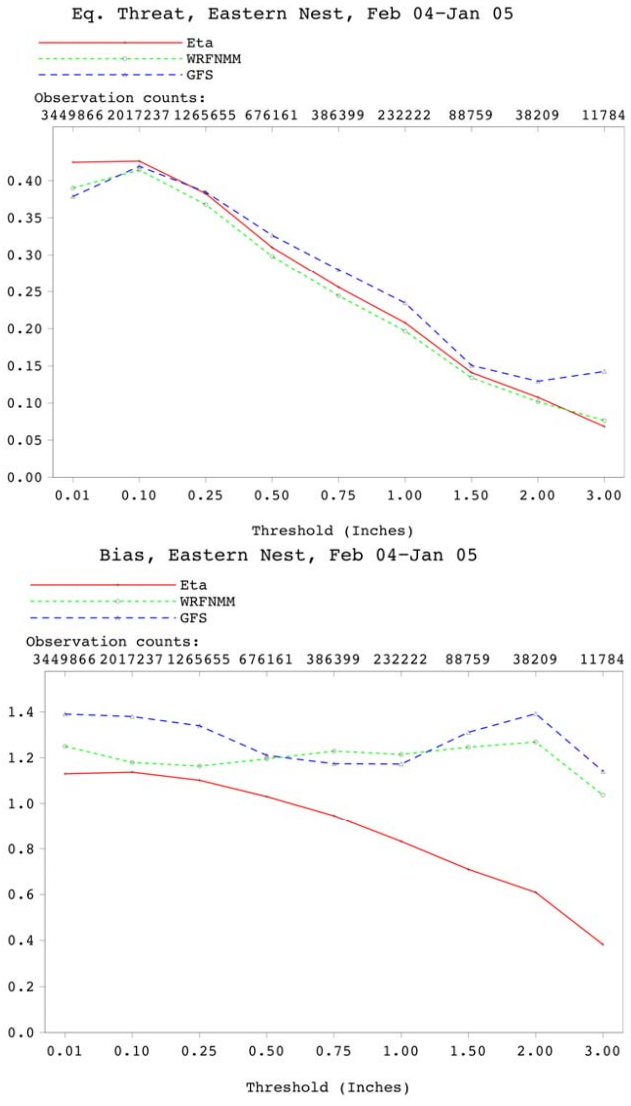

Fig. 3. 12-month precipitation equitable threat (upper panel) and bias scores (lower panel) for three NCEP operational models, "Eastern Nest", 18–42 h forecasts. See text for further detail.

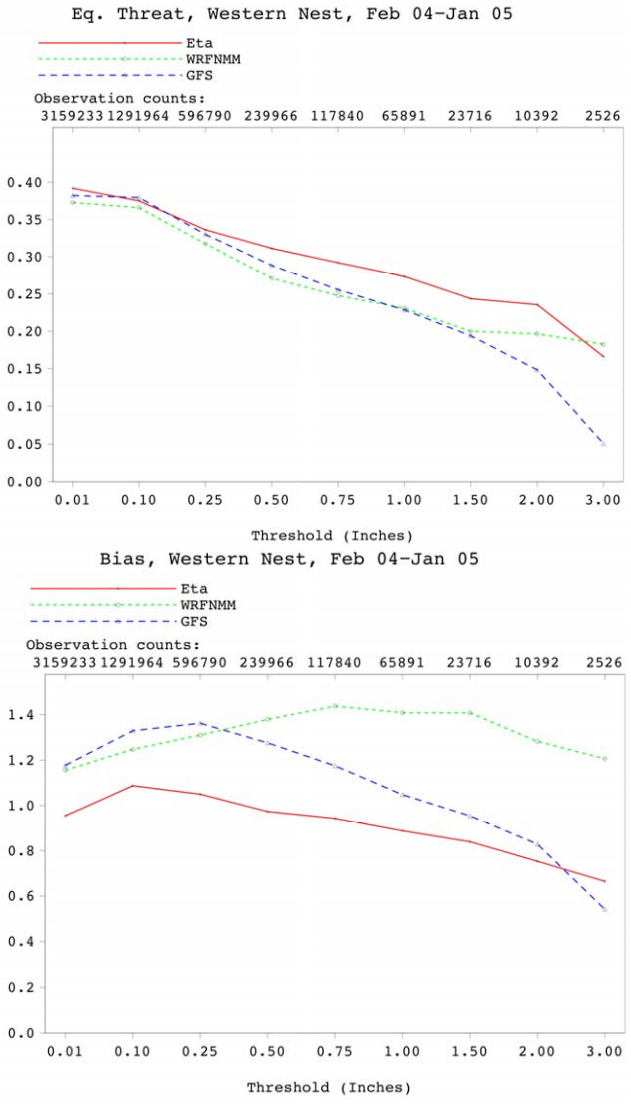

Fig. 4. 12-month precipitation equitable threat (upper panel) and bias scores (lower panel) for three NCEP operational models, "Western Nest", 6–30 h forecasts. See text for further detail.

Verification is performed on a 12 km grid. Model forecasts are remapped to the verification grid in a procedure in which precipitation is considered constant over the model grid-box. For more details on the precipitation verification system, maintained by Ying Lin, see http://www.emc.ncep.noaa.gov/mmb/ylin/pcpverif/scores/. The system is a component of and plots to be shown are generated by the NCEP Forecast Verification System (FVS), maintained by Keith Brill.

Equitable threat and bias scores for the 12-month period over the two domains are shown in Figs. 3 and 4, for the East, and for the West, respectively. Threat scores are shown in the upper panels of these figures and biases in their lower panels. In the East differences in threats between the models are not large; yet, the GFS threats are overall clearly still the highest, with the Eta threats being slightly above the NMM's. But there are large differences in biases: high biases of the GFS and the NMM are seen, compared to lower biases of the Eta, rather close to unity for low and medium intensity thresholds but then steadily decreasing toward the heavy precipitation end of the plot.

In the West, a considerable advantage of the Eta in threat scores is seen, with the Eta scoring the highest comfortably at all of the medium and high intensity thresholds from 0.5 up to and including 2 in/24 h. Differences between the GFS and the NMM threat scores are seen to be small, except for a clear advantage of the NMM at the two highest intensity thresholds. But once again there are large differences in biases, with the NMM bias scores being overall the highest,

those of the GFS being generally smaller, and the Eta ones being smaller still, and fairly accurate over most of the intensity range covered.

How much and in what way have these threat scores been affected by model biases is not obvious. Did the biases of around 1.2 help the GFS and the NMM in the East as "common wisdom" has it they should, and have the biases hurt the GFS when they reached values as high as about 1.4? In the West, did the Eta achieve these considerably higher threat scores because the NMM was hurt by its excessive biases at the medium to heavy intensity thresholds? Could it be that the reason that models for heavier precipitation did not achieve scores ranked the same as the resolution they used, as one might expect they should, is the impact of the bias on ETS? These are precisely the issues which the bias adjustment ought to help resolve.

Bias adjusted threat scores resulting from the present method are shown in Fig. 5, for the East, upper panel, and for the West, lower panel. In the East at the low intensity end of the plot, the biases of around 1.2 are indeed seen to have helped the Eta and the NMM ETS values, so that adjusted for bias these values have been reduced. In contrast, the very high bias of about 1.4 at this low end of the plot is seen to have been hurting the GFS, so that adjusted for bias its value has increased some. But at the high intensity end of the plot the same very high bias is seen to have been helpful to the GFS threats, as have of course also been the somewhat elevated biases of both the GFS and NMM throughout the medium and high intensity thresholds.

This difference in the benefit from the elevated bias at the lowest and at the higher intensity thresholds can be understood by a simple inspection of the impact of hits due to chance, FO/N in Eq. (2). At the low end of the thresholds monitored this term is large so that a relatively large expansion of the forecast area is needed for a given increase in hits above those due to chance, and in fact larger and larger as F increases. Thus, increasing the bias beyond unity ceases being beneficial to the increase in ETS at values that are not exceedingly high. As the threshold intensity increases FO/N decreases and eventually becomes negligible compared to even a very few hits, so that the benefit the ETS scores tend to have from extra hits resulting from inflated biases keeps being beneficial for very large biases.

In the West, with the bias of the NMM at the lowest threshold not higher than around 1.2, the inflated biases of the NMM are seen to have been of considerable help to its ETS values across all of the thresholds. The same holds for the GFS over lower thresholds albeit to a lesser degree. Thus, bias corrected ETS scores are seen to strongly suggest a higher placement accuracy of the Eta compared to both NMM and GFS across all of the thresholds.

As to the ranking of the models' ETS scores for heavier precipitation not being according to the resolution used, the bias correction shows that if the higher resolution was helpful to the models' placement accuracy, in case of the NMM

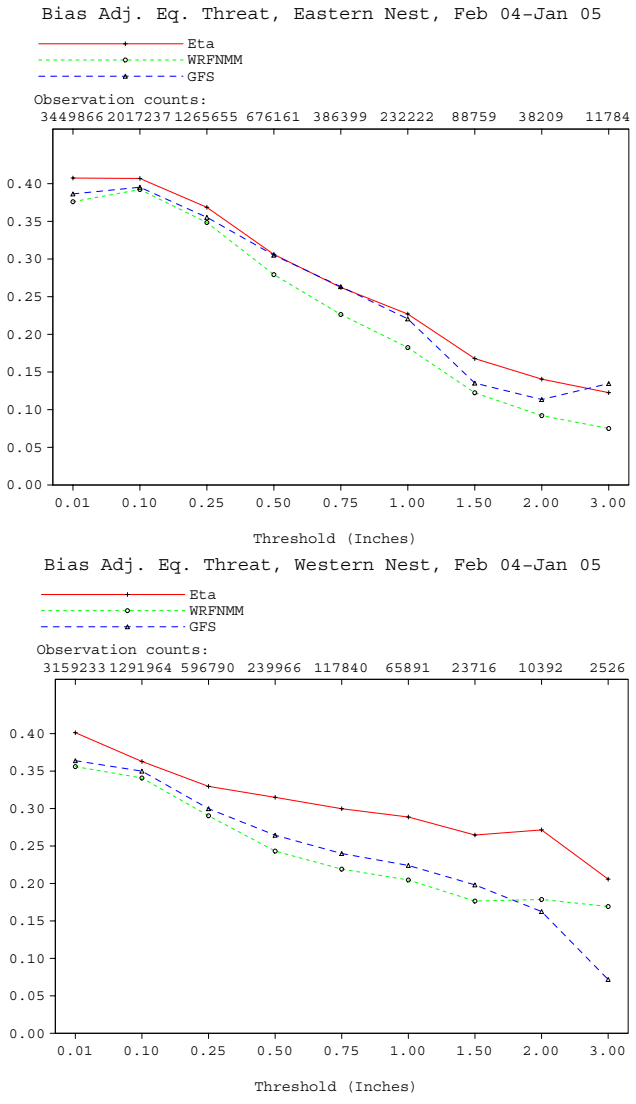

Fig. 5. Equitable threat scores as in the upper panels of Figs. 4 and 5, but adjusted to remove the effect of bias, using the dH/dA method. "Eastern Nest" upper panel, "Western Nest" lower panel.

there had to be other factor or factors in place and not the model's bias that have more than offset the resolution benefit the model may have had.

4 Summary

It is pointed out that the widespread use of the ETS and bias scores "as a measure of forecast accuracy" is misleading, since higher ETS scores will normally result from biases inflated beyond unity. A method is proposed to adjust or correct ETS scores so as to remove the impact of bias, and thus arrive at a measure that reflects the model's accuracy in *placing* precipitation. It is suggested that this bias adjusted ETS along with bias should be a much more useful information

than the standard ETS and bias that for years many authors have been using to assess the accuracy of precipitation forecasts.

Acknowledgements. Joseph Schaefer pointed out to me the desirability of "Unbiasing the CSI" (subject line of his e-mail, 2002). Extensive discussions with Mike Baldwin, and Keith Brill, were extremely helpful in my arriving at the present state of the manuscript. Additional input from Tom Black, Beth Ebert, Bob Glahn, Tom Hamill, Dušan Jović, Wes Junker, Ying Lin, Eric Rogers, Joe Schaefer, and the Advances in Geosciences reviewers helped in a variety of ways. Andrei Mesinger assisted in overcoming an obstacle that at one point held my progress with the proposed method.

Edited by: S. C. Michaelides
Reviewed by: two anonymous referees

References

Baldwin, M. E. and Kain, J. S.: Sensitivity of several performance measures to displacement error, bias, and event frequency, Wea. Forecasting, 21, 636–648, 2006.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, Mon. Weather Rev., 134, 1772–1784, 2006.

Ebert, E. E. and McBride, J. L.: Verification of precipitation in weather systems: Determination of systematic errors, J. Hydrol., 239, 179–202, 2000.

GCWM Branch, EMC: The GFS Atmospheric Model. NCEP Office Note 442, 14 pp., available at: http://wwwt.emc.ncep.noaa.gov/officenotes, 2003.

Gilbert, G. F.: Finley's tornado predictions, Amer. Meteorol. J., 1, 166–172, 1884.

Hamill, T. M.: Hypothesis tests for evaluating numerical precipitation forecasts, Wea. Forecasting, 14, 155–167, 1999.

Janjić, Z. I.: A nonhydrostatic model based on a new approach, Meteorol. Atmos. Phys., 82, 271–285, 2003.

Mason, I.: Dependence of the critical success index on sample climate and threshold probability, Aust. Meteorol. Mag., 37, 75–81, 1989.

Mesinger, F. and Brill, K.: Bias normalized precipitation scores. Preprints, 17th Conf. on Probability and Statistics, Seattle, WA, Amer. Meteorol. Soc., CD-ROM, J12.6, 2004.

Pielke, R. A., Sr.: Mesoscale Meteorological Modeling, 2 ed., Academic Press, 676 pp., 2002.

Schaefer, J. T.: The critical success index as an indicator of warning skill, Wea. Forecasting, 5, 570–575, 1990.

Shuman, F. G.: A modified threat score and a measure of placement error. NOAA/NWS Office Note 210, 13 pp., available at: http://wwwt.emc.ncep.noaa.gov/officenotes, 1980.