J12.6 BIAS NORMALIZED PRECIPITATION SCORES

Fedor Mesinger* and Keith Brill^

* NCEP/EMC and UCAR, ^ NCEP/CPC, Camp Springs, Maryland

Abstract

The purpose of the equitable threat or various related scores is to provide information on model's accuracy in placing precipitation above a given threshold. Yet, they do not quite manage to achieve this because of their dependence on bias, so that a subjective assessment of how and how much these scores might have been affected by the model bias is customary. Conversely and as an opportunity for improving scores in a manner that can be considered as ethically questionable, common wisdom has it that a bias somewhat greater than one is profitable.

It is shown that a more satisfactory state of affairs can be arrived at. A simple assumption of the increase of hits per unit increase in bias being proportional to the yet unhit area enables calculation of the number of hits normalized to a perfect bias. Thus, normalization of the equitable threat and related scores to perfect bias is possible. Assumption of the odds ratio being independent of bias can be used to the same end. Examples of the resulting bias normalized equitable threat scores of several operational NCEP models are presented.

1. Introduction

The purpose of the threat scores, standard or equitable, is to assess the skill of a model in placing its forecasts of an event, say precipitation above a given threshold. There are quite a few other statistical quantities aiming for roughly the same objective, but equitable threat may well be the most popular. The problem of the skill assessment in a situation where there is a forecast of an event, that can either occur or not occur, is of course quite general, common to many fields. Note Murphy (1996) for an entertaining account of the early weather prediction efforts of more than a century ago, with various quantities reintroduced and renamed later, some more than once. The reason why threat and equitable threat score (or, Gilbert score, Schaefer 1990) are in meteorology more popular than some of the other measures is that threat and equitable threat emphasize skill in forecasting the occurrence of the event more than they do the skill in

* *Corresponding author address*: Fedor Mesinger, NCEP Environmental Modeling Center, 5200 Auth Road, Room 207, Camp Springs, MD 20746-4304; e-mail: fedor.mesinger@noaa.gov

forecasting the non-occurrence of the event. Successful forecast of a heavy rain event, say more than an inch in 24 hours, certainly results in more respect for the forecaster than does a successful forecast that this heavy rain will not occur, few will disagree.

Over the years, a persistent problem in using the equitable threat score, or some of the alternative measures, was their dependence on bias. Thus, typically, when displaying an equitable threat score plot, the corresponding bias plot would be displayed as well, and would have to be taken into account. Many examples to that effect could be referred to, for a recent one, see, e.g., Ebert et al. (2003). Assessing to what extent model's placement forecasts were accurate, one would inspect not only the threat score plot but its bias plot as well. Or, comparing the relative placement performance of two models, their relative bias performance would have to be taken into account. The subjective nature of this procedure clearly diminishes the value of the conclusions made.

Efforts to alleviate this annoying situation seem so far to have been less than entirely successful. If we do have a contingency table with four outcomes, forecasts of yes and no, and occurrences of yes and no -- which is not always the case -- "hedging" is available (Stephenson 2000). For example, with bias greater than one, one can randomly remove forecasts so as to obtain a bias of one. Given that model forecasts are not random, this clearly is not a very attractive scheme.

Hamill (1999) presents an illustration how an equitable threat of a forecast with perfect bias would increase, up to a point, if contours were to be relabeled toward an inflated bias. This underscores a general understanding that for optimum threats a bias somewhat above one is needed, which is clearly a goal deserving less than a complete respect. Hamill's example suggests that, conversely, with a bias different from one and a contour plot available, relabeling isolines toward elimination of bias is a possible technique. The same idea can be pursued clearly also with no contours, by way of relableling the forecast values. But there are obvious problems: when is the relabeling to be done, for each forecast, or after collecting a sample, e.g., a month's worth, of F, H, O (forecast, hits, and observed) values for a set of precipitation categories? Presumably the latter, followed by interpolation to arrive at bias equal to one. How does one interpolate to arrive at hits corresponding to bias of one is not obvious. Besides, interpolation from the lowest category, or categories, is or may not be possible when the bias is less than one.

There are other efforts to verify precipitation using approaches that do not suffer from the bias problem, or alleviate it in a way that is related to the idea of Hamill (e.g., Ebert and McBride 2000; Atger 2001). But it seems to us that, nevertheless, the simplicity of the threat or equitable threat, and the widespread use they enjoy, make the correction of these for the impact of bias a

worthwhile goal. We shall look into two possibilities we have pursued to achieve that. The two sections to follow will each be devoted to presenting one of our two methods. This will be followed by examples of the results, and by brief overall comments.

2. The dH/dF method

We want to correct or adjust the threat or equitable threat for the impact of bias, or, bias normalize one or the other of these scores. In other words, we wish to obtain their values corresponding to bias of one. To that end, one needs an assumption how should the number (or, area) of hits, H, be expected to increase with an increase in the number (area) of forecast events, F. For ease of analysis, let us assume that these are continuous quantities, as well as the area of observed events, O. Results can be dicretized later.

Whatever the model's inherent skill, one should expect that it is easier for a model to score a hit, or increase its hits area, when the availability of events that yet remain to be hit is greater. Thus, we assume that the increase in hits area per unit increase in F is proportional to O-H,

$$\frac{dH}{dF} = a(O - H), \quad a = const. \tag{1}$$

Solution of (1) is

$$H(F) = be^{-aF} + O, \quad b = const. \tag{2}$$

Since we have $H = 0$ for $F = 0$, we obtain

$$b = -O.$$

Thus,

$$H(F) = O\left(1 - e^{-aF}\right). \tag{3}$$

Solving for a, this gives

$$a = -\frac{1}{F}\ln\left[1 - \frac{H(F)}{O}\right]. \tag{4}$$

We can calculate a from a known set of values of H, F, and O, insert it into (3), and use it to obtain the value of H adjusted for bias. Denote this known set of values by H_b, F_b, and O (note that O is considered given; we are only aiming to determine how H should change with F). Thus,

$$a = -\frac{1}{F_b}\ln\left(1 - \frac{H_b}{O}\right). \tag{5}$$

Inserting this into (3) results in

$$H(F) = O\left[1 - \left(\frac{O - H_b}{O}\right)^{\frac{F}{F_b}}\right]. \tag{6}$$

This is the desired dependence of H on F. We just need the specific value of H corresponding to bias being 1, that is, for $F = O$; let us denote it as H_a (H adjusted)

$$H_a = O\left[1 - \left(\frac{O - H_b}{O}\right)^{\frac{O}{F_b}}\right]. \tag{7}$$

There is no need any more to differentiate between specific values of H and F, H_b and F_b above, used to determine a in (3), and H as a function of F. We can thus omit the subscripts b above, and write

$$H_a = O\left[1 - \left(\frac{O - H}{O}\right)^{\frac{O}{F}}\right]. \tag{8}$$

This permits one to calculate the threat or equitable threat score adjusted for bias, that is, its value corresponding to $F = O$. When doing this, $F = O$ should of course be used for hits due to chance, that is, use $E(H) = FO/N = OO/N$, N here being the total number of events.

For a simple numerical example, consider a very rare event, observed at 100 points out of a total of 300 x 200 = 60,000; with $F = 50$ and $H = 20$. In this case correction for chance events is very small, $FO/N = 0.08333...$ One obtains for the equitable threat, T_e

$$T_e = 0.1533...$$

(8) results in

$$H_a = 36.$$

Note that with the first 50 forecast points resulting in 20 hits, the next 50, according to (8), results in only 16 additional hits. With only 80 event-points left, as opposed to 100 available with $F = O$, it is harder for the model to achieve new hits.

To obtain bias adjusted equitable threat, T_{ea}, we use H_a, and $FO/N = OO/N = 0.1666...$, resulting in

$$T_{ea} = 0.2187...$$

an increased value compared to 0.1533.

3. The odds ratio preserving method

Stephenson (2000) has argued that the "odds ratio", widely used in medical studies, is a powerful measure for verification of categorical forecasts, listing a number if its useful properties. Note its recent use for precipitation by Goeber and Milton (2002). The "odds", or the "risk" of an event, is the ratio of the probability of an event occurring to the probability of the event not occurring. Stephenson writes his odds ratio definition in terms of non-marginal contingency table elements, hits, false alarms, misses, and correct forecasts of non-occurrence, denoted by a, b, c and d, respectively. For ease of reference relative to the notation H, F, and O of the preceding section, we are displaying an example of a possible pattern in Fig. 1.

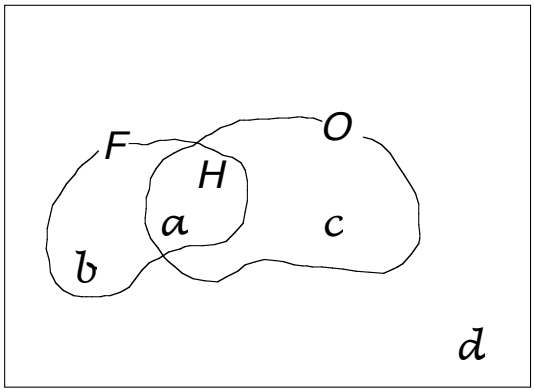

Fig. 1. Schematic of the relationship between the forecast, *F*, hits, *H*, and observed, *O*, values, and the non-marginal contingency table elements, *a*, *b*, *c*, and *d*.

The odds ratio, Γ, as stated by Stephenson (2000), is "the ratio of the odds of making a hit given that the event occurred to the odds of making a false alarm given that the event failed to occur". The former of the odds' is defined as

$$\left[\frac{a}{a+c}\right] \Big/ \left[1 - \frac{a}{a+c}\right], \tag{9}$$

and the latter as

$$\frac{b}{b+d}\Big/\left(1-\frac{b}{b+d}\right). \tag{10}$$

Thus, the odds ratio is

$$\theta=\frac{\frac{a}{a+c}\Big/\left(1-\frac{a}{a+c}\right)}{\frac{b}{b+d}\Big/\left(1-\frac{b}{b+d}\right)}, \tag{11}$$

which results in

$$\theta=\frac{ad}{bc}. \tag{12}$$

The numerator of (9), $a/(a+c)$, is referred to as the hit rate, or probability of detection, and that of (10), $b/(b+d)$, the false alarm rate. In the H, F, O and N notation of the preceding section, they are equal to H/O, and $(F-H)/(N-O)$, respectively. Thus, after rearrangement, (11) can also be written as

$$\theta=\frac{H(N-O-F+H)}{(O-H)(F-H)}. \tag{13}$$

To adjust to bias equal to unity under the assumption of the preservation of odds ratio, first θ is computed using (13). Then, with F set to O, (13) is solved for the adjusted value of H, H_a, giving

$$H_a=O+\frac{N}{2(\theta-1)}-\left\langle\left[O+\frac{N}{2(\theta-1)}\right]^2-\frac{\theta}{\theta-1}O^2\right\rangle^{1/2} \tag{14}$$

The adjusted set of F, H, O values becomes $F=O$, H_a, O. These values are used to compute odds ratio preserving bias adjusted scores. This method of adjustment assures that both the forecast and the adjusted forecast lie on the same relative operating characteristic (ROC) curve as parameterized in terms of the odds ratio (Stephenson 2000). A ROC curve is graphed by plotting the probability of detection against the false alarm rate for a collection of forecast verifications.

4. Examples: 12 months of three model scores

For examples of the impact, we shall show precipitation scores of three NCEP operational NWP models over two regions of the contiguous United States (ConUS). One is the Global Forecast System (GFS), containing the NCEP operational global model, referred to also as GFS. The other is the short range (out to 3.5 days) Eta model, run at 12 km horizontal resolution. The Eta model is a limited area model, obtaining its lateral boundary conditions from the GFS run of 6 h earlier. As of summer 2002, a still higher resolution model, NMM (Nonhydrostatic Mesoscale Model) is run over six "high resolution windows" covering the 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 is run at 8 km horizontal resolution.

Comparison of the relative performance of these models is of interest for numerous reasons. While there are a number of verification efforts in place, precipitation equitable threat scores, as

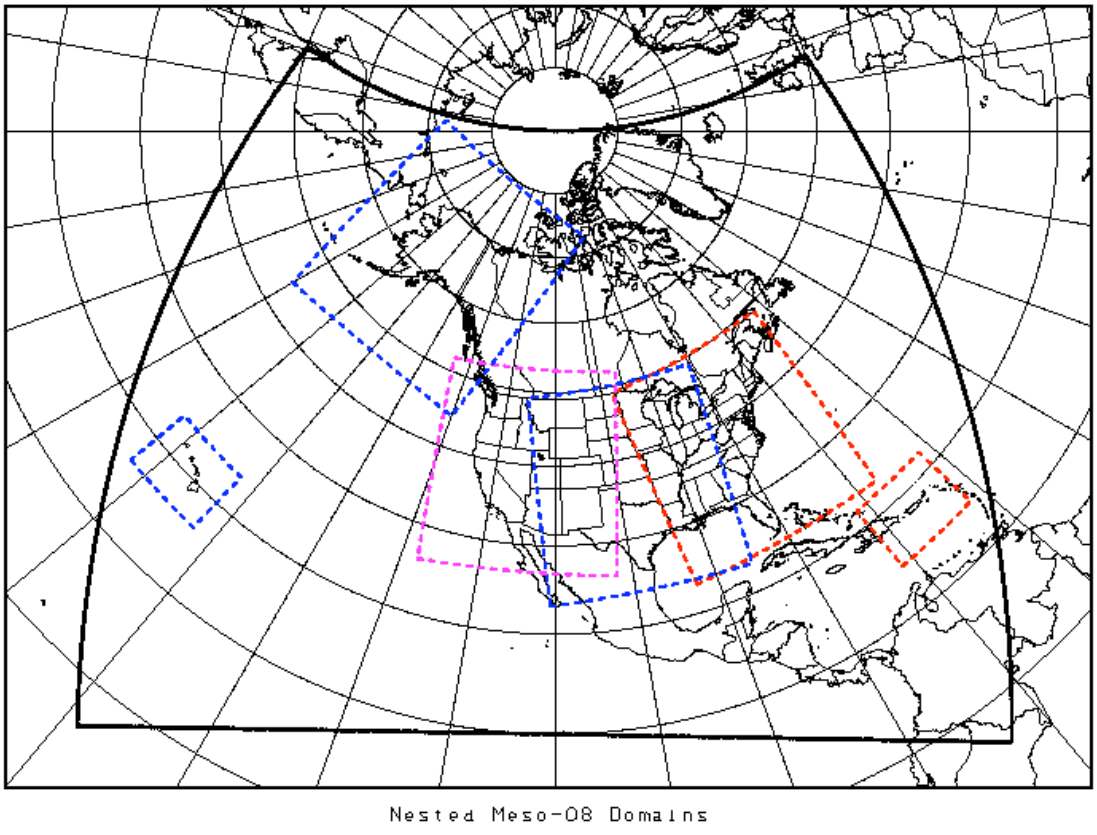

Fig. 2. Domains of the Eta 12 km operational model, heavy black line, and of the "high resolution windows" of the nested NMM model (Nonhydrostatic Mesoscale Model), dashed color lines.

described earlier in combination with the bias scores, are certainly among those of most interest if not in fact *the* statistic of most interest (e.g., Mesinger 1996). QPF scores over the three ConUS NMM domains have become available within the NCEP Forecast Verification System (FVS) as of September 2002. Verification on these NMM domains is performed on a 12 km verification grid. Forecasts of models with resolution or native grid different from the verification 12 km grid are remapped to the verification grid. In this procedure precipitation is considered constant over the model grid-box, and is numerically integrated to the verification grid, with a desired accuracy.

In the Eta and NMM performance comparison issues of interest include those of the expected benefit from the increased resolution of the NMM compared to the Eta, and perhaps from its nonhydrostatic feature. The increase in model resolution is believed to be an important factor for improvement of forecasts of intense precipitation (e.g., Buizza et al 1997). Besides, in the NMM the eta coordinate of the Eta has been abandoned in favor of the traditional terrain-following (sigma) coordinate. This was based primarily on problems encountered with downslope windstorms using the eta (Janjic 2002, and references therein), and perhaps on a general expectation that this is a sign of difficulties to be expected with the eta as the resolution is increased. But just the opposite could also be expected (Mesinger, this CD). With these issues at hand, comparison of the relative performance of the Eta and the NMM over the Eastern Nest ("East"), where there is little influence of topography, and over the Western Nest ("West"), where the topography is dominant, could offer significant clues as to what the dominant impacts actually are.

Equitable threat and bias scores for the first 12 months of the availability of scores over the NMM domains are shown in Figs. 3 and 4, for the East, and for the West, respectively. In the East, threats of the Eta and the NMM are just about the same, if anything, those of the Eta are slightly better. Biases of the two are very nearly the same as well. Thus, no benefit from higher resolution, 8 vs 12 km, is seen in QPF scores. There could of course exist handicaps that the NMM was facing relative to the Eta which might have prevented it from generating overall better scores; it is not within the scope of the present paper to go into such issues.

The GFS over the East, shows threats much better that the Eta and the NMM for higher intensity categories, and an inferior threat score for the lowest category of 0.01 inch/24 h. But it has a much higher bias than the Eta and the NMM for higher categories, and a considerably higher bias for the three lowest categories. While the huge threat advantage for higher categories could hardly have resulted from higher bias alone, the impact of bias at the lowest category is not easy to tell.

Over the West, where one would expect the higher resolution topography of the 8-km model

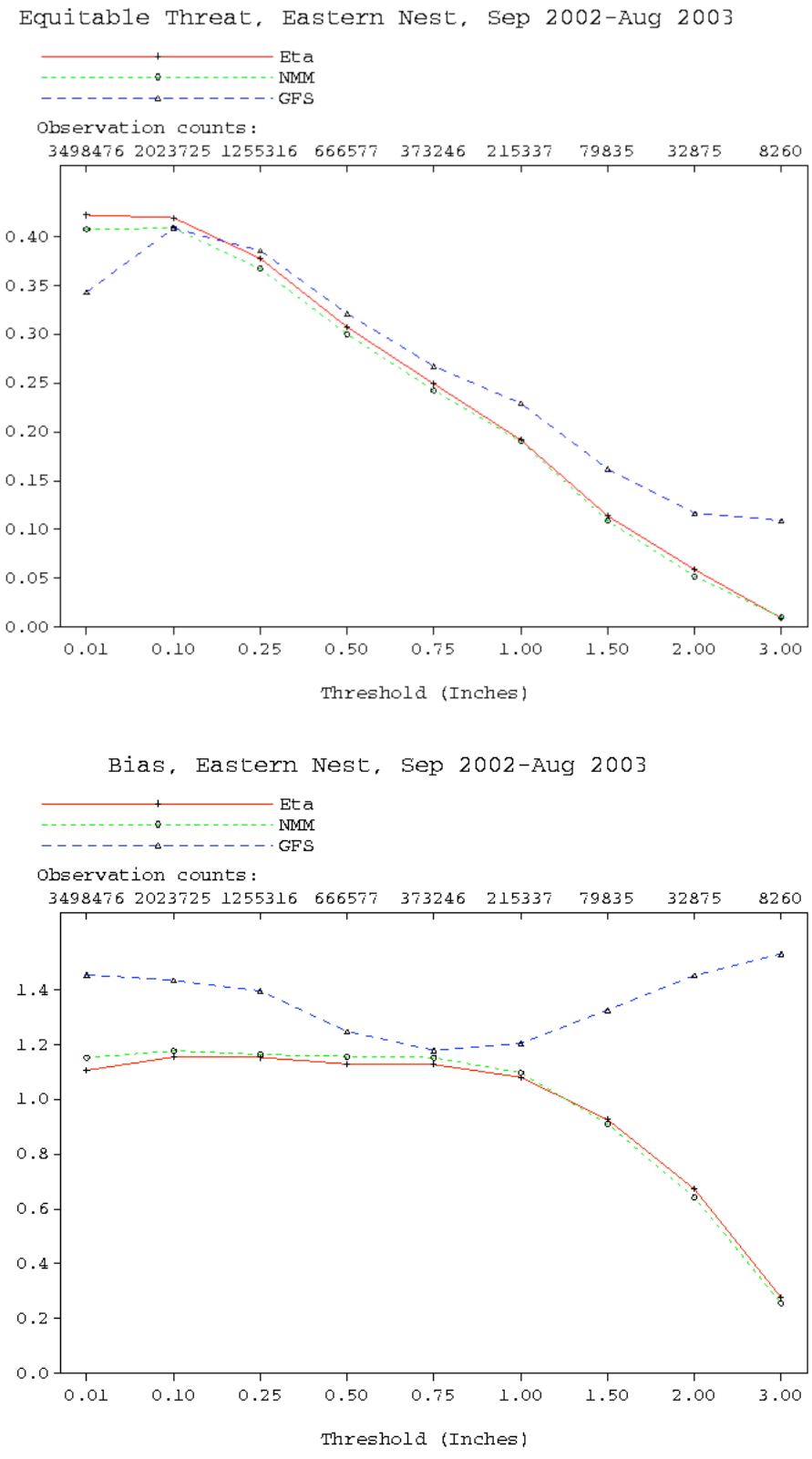

Fig. 3. 12-month precipitation equitable threat (upper panel) and bias scores (lower panel) for three NCEP operational models, "Eastern Nest", 18-42 h forecasts. See text for further detail.

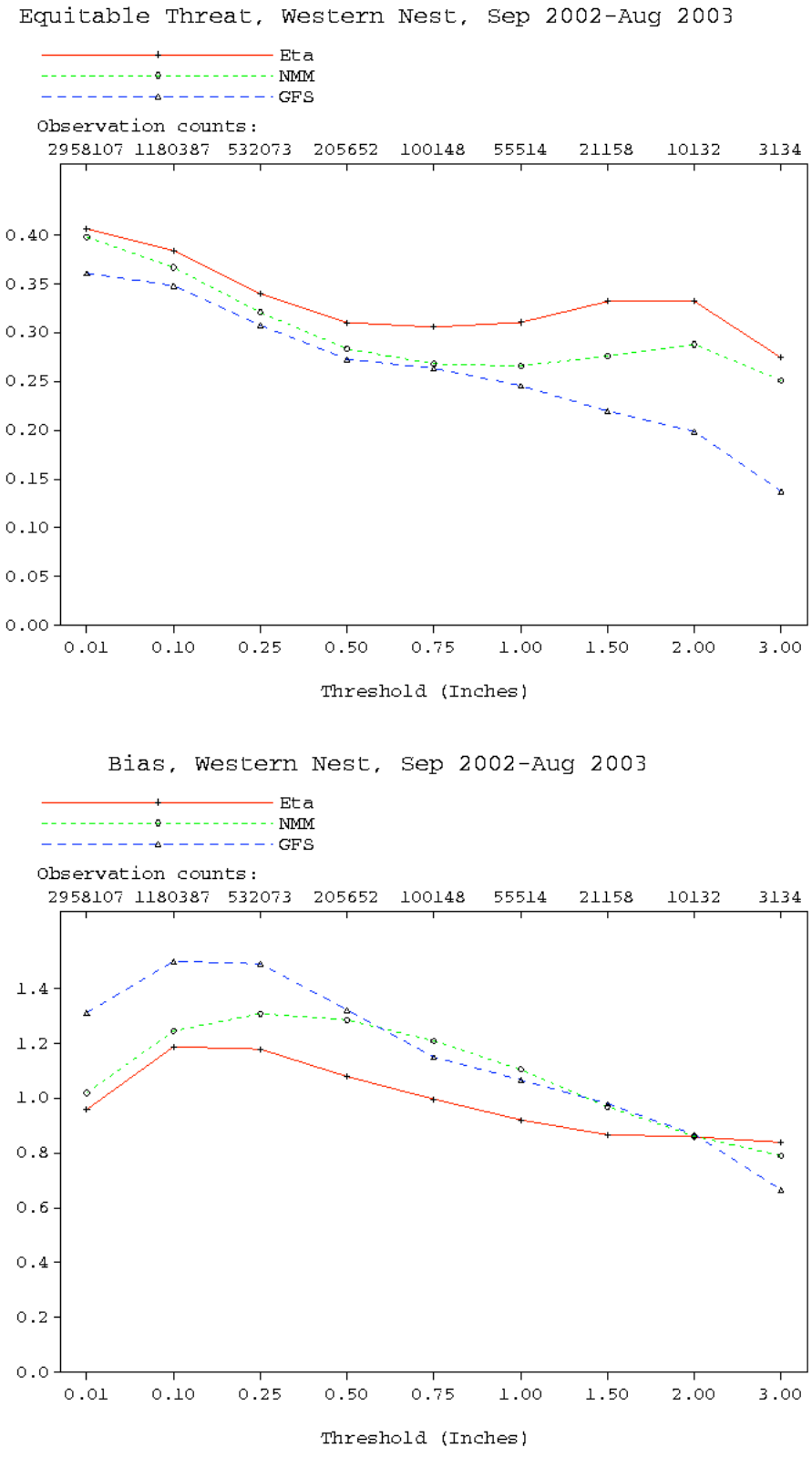

Fig. 4. 12-month precipitation equitable threat (upper panel) and bias scores (lower panel) for three NCEP operational models, "Western Nest", 6-30 h forecasts. See text for further detail.

to be of a particular advantage, the Eta 12-km model is seen to have equitable threats substantially better than the 8-km NMM. The Eta advantage was particularly striking in situations of intense precipitation over the western United States in November and December 2002 (http://wwwt.emc.ncep.noaa.gov/mmb/ylin/pcpverif/scores/), with five events of HPC analyzed precipitation over 4 inches/24 h, and two of these with precipitation over 6 inches/ 24 h. At these events, the highest precipitation monitored of 3 inches/24 h, would typically be analyzed over individual and separated mountain ranges, Cascades, Coast Ranges, and Sierras, with elongated and clearly topographically defined patterns that are generally considered an extraordinary QPF challenge. At the same time, both high resolution models are having threats better than the GFS, the Eta much better. But with the elevated bias of the NMM at medium categories, one may wonder if the disadvantage of the NMM compared to the Eta is at least partly a result of its bias problem. Just as well, one may wonder if the inferior threats of the GFS at the lowest categories are the result of its considerably increased bias.

These are precisely the issues which bias normalization ought to help resolve. This being now available within the NCEP FVS, we have generated bias normalized equitable threats for the two domains using one and the other of the two methods. In Fig. 5 we are displaying the threat plots for the East, those of the upper panel of Fig. 3, bias normalized using the dH/dF method, upper panel, and using the odds ratio method, lower panel. Bias normalized scores for the West, those of the upper panel of Fig. 4, are shown in Fig. 6.

The two sets of bias normalized threats for the East, Fig. 5, are seen to be very similar, except at the lowest category. With both methods, with the equitable threat penalty for the high bias of the GFS at the lowest categories removed, the GFS is seen to be uniformly superior to the two mesoscale models across all categories. At the lowest category, the dH/dF method is seen to reward the GFS considerably more than the odds ratio method. The reason for this is that with the probability of detection, H/O, as high as that of the GFS at this lowest category, 0.92, the basic assumption of the dH/dF method, (1), distributes most of the hits for values of F/O less than one, so that the reduction in hits resulting from the reduction of bias to one is considerably smaller than it is with the odds ratio method. In case one feels that the bias normalized threat the GFS was accorded to by the dH/dF method at 0.01 inch/24 h is too large, one might find some comfort in the fact that the GFS' standard threat at this category is still considerably greater, about 0.60.

The same two plots but for the West, Fig. 6, are again quite similar, except at the lowest category. At that category the dH/dF method once more rewards the GFS more than the odds ratio method. But the results show that, except at this lowest category, equitable threats of the

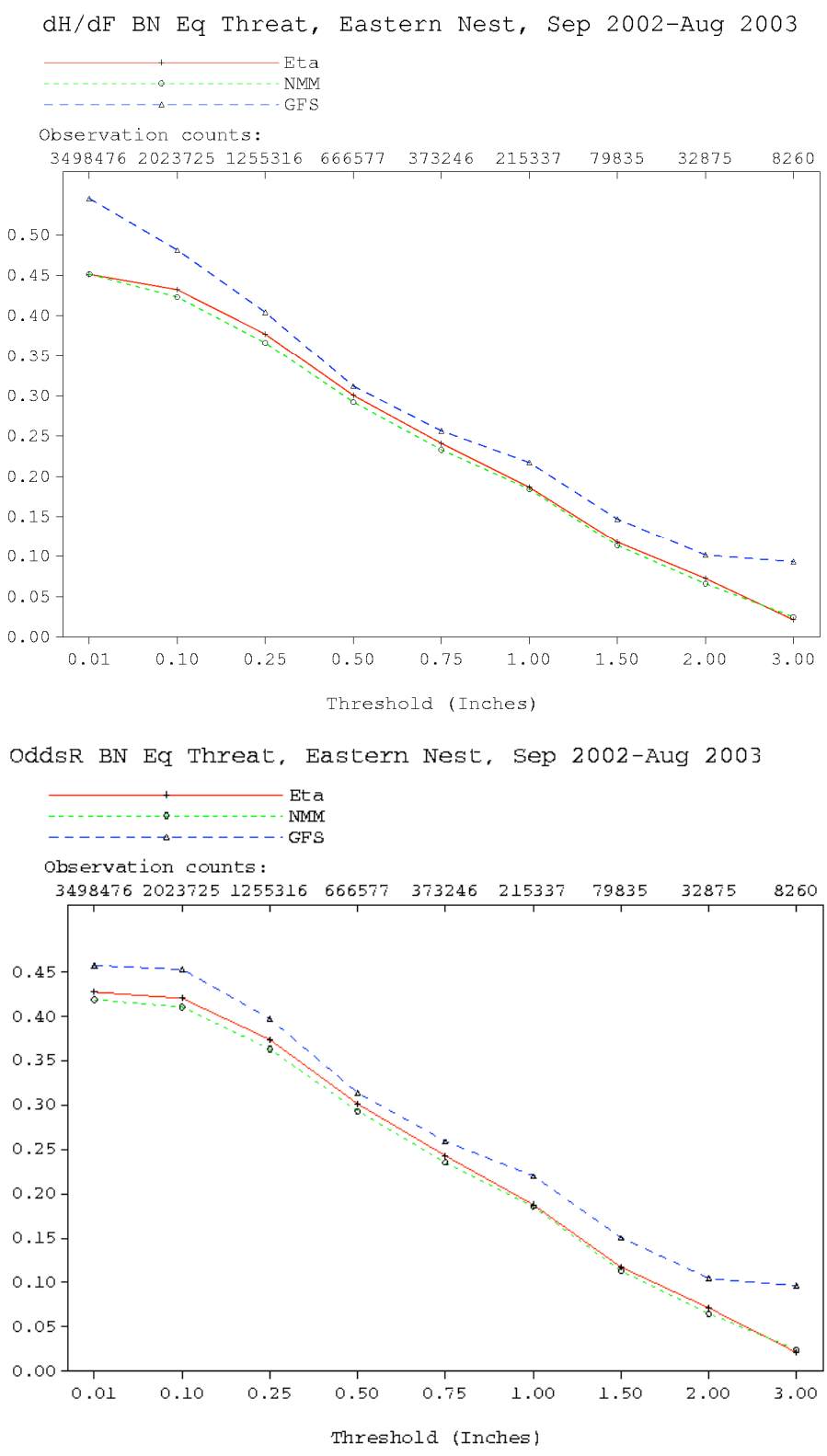

Fig. 5. Equitable threat scores as in the upper panel of Fig. 3, but normalized to remove the effect of bias using the "dH/dF method", upper panel, and using the odds ratio method, lower panel.

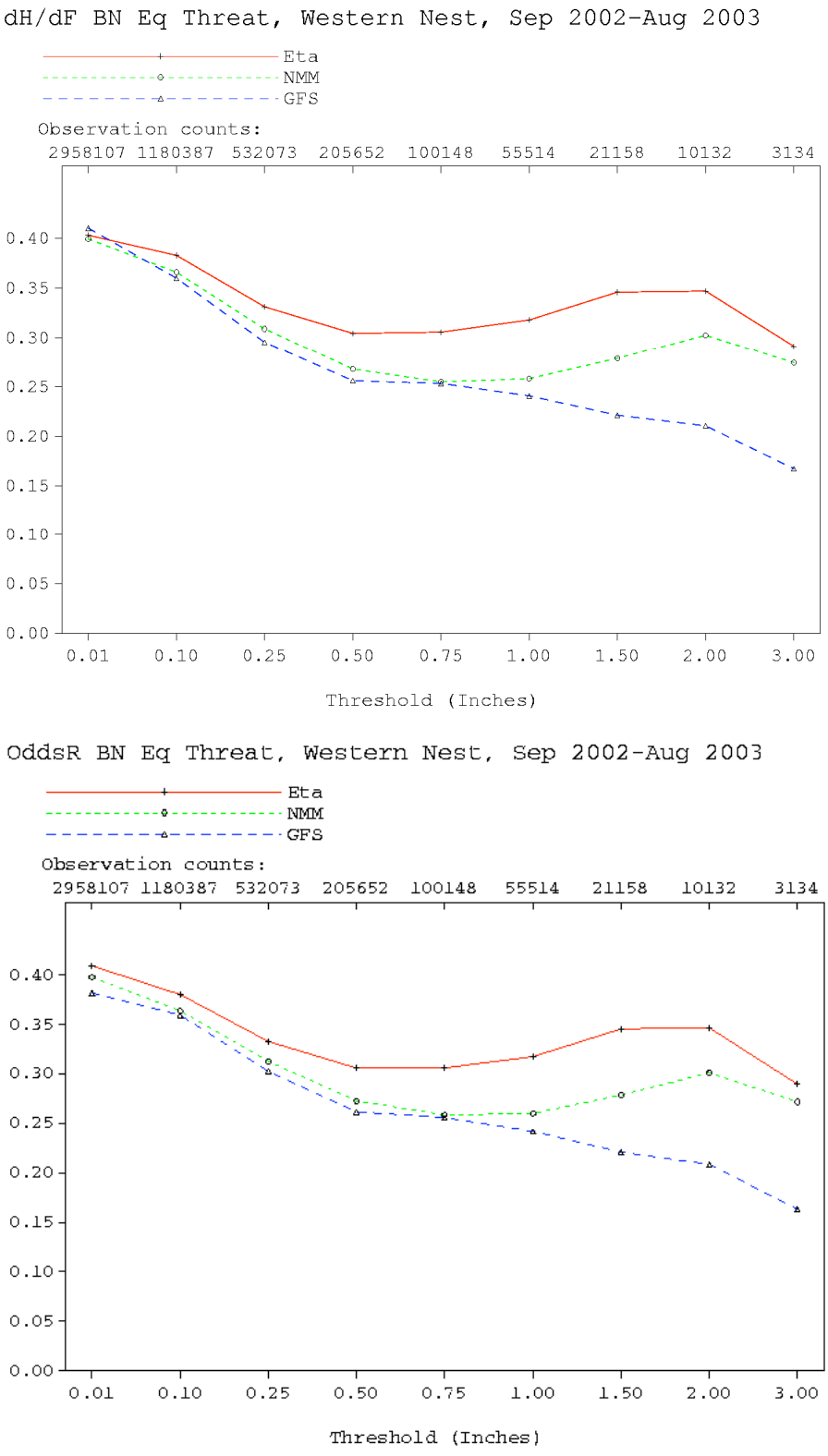

Fig. 6. Equitable threat scores as in the upper panel of Fig. 4, but normalized to remove the effect of bias using the "dH/dF method", upper panel, and using the odds ratio method, lower panel.

GFS and those of the NMM have not been depressed compared to those of the Eta due to the higher biases of these models. Quite to the contrary, it is seen that the increased biases of the GFS and the NMM of about 1.2 at medium categories, have been helping these models to achieve higher equitable threats, just as we said earlier "common wisdom" has it it should happen. As a result, the overall advantage of the Eta over the other two models in the West is seen to be higher when adjusted for bias than it seemed to be on account of the equitable threats alone, the upper panel of Fig. 4.

5. Discussion

To adjust threat or equitable threat scores to bias so as to arrive at scores corresponding to bias of one, an assumption is needed. The assumption of the dH/dF method, (1), is straightforward, and its single free parameter, assumed constant, can be determined from the available F, H, O values. The integration constant, b, is determined from the obvious condition that if F were equal to zero, H must be zero as well. Thus, what occurs as a result is a score based on hits resulting from interpolation, or extrapolation, of the obtained function $H(F)$ to the value of $F = O$.

If interpolation or extrapolation is performed to a relatively distant value of F, one might be uncomfortable about the outcome. We have already expressed concern about the GFS value obtained at the lowest category in the upper panel of Fig. 5. A possible attitude is to note that threat scores are really meant to emphasize hits, correct forecasts of an event, and not so much correct forecasts of no event, and that correct forecasts of no event played a large role at this category. Events observed to all elements ratio, O/N, at this category, was about 0.4. Yet another option is to try to refine the dH/dF scheme, by improving on the $a = const$ assumption. Note that as $F \prod N$ hits should approach O, a property that the odds ratio scheme has but the dH/dF scheme does not. But there likely are considerations that should have higher priority in attempting to improve on the $a = const$ assumption if this indeed were desirable, such as perhaps some based on actual model performance data.

The symmetry of the odds ratio scheme, (12), is esthetically appealing. Yet, the scheme could also be criticized for the same reason, in that it fails to emphasize hits compared to the correct forecasts of no event. Note that the numerators of (9) and (10) are used as axes to plot the relative operating characteristic (ROC) curve of Mason, a performance measure that appears to be gaining in popularity (Stephenson 2000; Atger 2001). According to Stephenson (2000), "The odds ratio is almost invariant with decision threshold" which supports the assumption of the conservation of odds ratio as a legitimate method of bias correction. At the same time, this is

suggestive of the method representing an approximation to or a variation of the Hamill method of contour relabeling, the appealing side being that the actual performance of the model is taken advantage of, as opposed to working with a hypothetical model behavior. Baldwin and Kain (2004), after examining a number of performance measures, find that the odds ratio skill score, $(\square 1)/(\square + 1)$, of measures examined, "is the least sensitive to bias error and the event frequency", which also would seem to support the reasonableness of the assumption of the odds ratio scheme.

But whatever the preference of a potential user, or further developments concerning the issues raised, we feel that the basic question is one of whether the set of bias normalized equitable threat and bias scores gives better model QPF information than the set of standard equitable threat and bias scores, that we tend to use today. We are convinced that it is, since the bias normalized threats can hardly fail to give information on model precipitation placement errors much less influenced by model bias than the standard threats are. Various issues here only touched upon we expect will become clearer as a result of additional work, some of which we hope to do ourselves.

Acknowledgements. Joseph Schaefer, of the Storm Prediction Center, pointed out the desirability of "Unbiasing the CSI" (subject line of his e-mail, 2002), and has suggested the term "bias normalization". Extensive discussions with Mike Baldwin, of the National Severe Storms Laboratory, have been most helpful in advancing our understanding of the behavior of the dH/dF scheme; Mike Baldwin has also introduced us to some of the related work in the area. Eric Rogers, of the Environmental Modeling Center (EMC), generated the plot we show in Fig. 2. Ying Lin, also of EMC, is maintaining the EMC's precipitation verification system, a component of the Climate Prediction Center's (CPC) forecast verification system, maintained by Keith Brill. These systems were used to generate the plots shown in Figs. 3 to 6.

References

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Proc. Geophys.*, **8**, 401-417.

Baldwin, M. E., and J. S. Kain, 2004: Examining the sensitivity of various performance measures. *17th Conf. on Probability and Statistic in the Atmospheric Sciences*, 84th AMS Annual Meeting, Seattle, WA, January 2004 (this CD-ROM).

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1997: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935-1960.

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

Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc*., **84**, 481–492.

Goeber, M., and S. F. Milton, 2002: Verifying precipitation events forecast by the mesoscale model. *NWP Gazette*, March 2002, 9-11. [Available at www.metoffice.com/research/nwp/publications/nwp_gazette/index.html.]

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

Janjic, Z. I., 2002: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys*., **82**, 271-285.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional Model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc*., **77**, 2637-2649; Corrigendum, **78**, 506.

Murphy, A. H., 1996: The Finley Affair: A signal event in the history of forecast verification. *Wea. Forecasting*, **11**, 3–20.

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

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting*, **15**, 221-232.