# 2.9 EXAMINING THE SENSITIVITY OF VARIOUS PERFORMANCE MEASURES

Michael E. Baldwin*[1] and John S. Kain[2]
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK.
1 Also affiliated with NOAA/NSSL and NOAA/SPC
2 Also affiliated with NOAA/NSSL

## 1. INTRODUCTION

Performance measures, such as the threat score, are widely used as summary measures of forecast quality. Depending upon the type of forecast being issued, whether continuous or categorical, probabilistic or deterministic, there are a variety of measures from which to choose. For example, the Environmental Modeling Center (EMC), a part of the National Centers for Environmental Prediction (NCEP), primarily uses equitable threat and bias scores to quantify the performance of precipitation forecasts from numerical guidance. Forecasters, managers, and users of forecast information require verification information in order to answer several questions: "How are the forecasts performing? How can future forecasts be improved? How does one forecast system compare to another?" The validity of verification information that is used to answer such questions depends upon the characteristics and sensitivies of the scores used.

Typical performance measures provide information on a single aspect of forecast quality: forecast *accuracy*. Accuracy is related to how well the forecasts correspond to the observed truth. Accuracy is one of the many aspects of forecast quality (Murphy 1993) provided by the joint distribution of forecasts and observations that must be analyzed in order to obtain a complete diagnosis of verification information. However, the distributions-oriented approach is rarely used in practice due to the complexity and high dimensionality of the joint distribution of forecasts and observations. Instead, accuracy measures are often used as a substitute to provide an overall summary of the forecast quality (called the measures-oriented approach by Brooks and Doswell 1996).

When selecting an accuracy measure, one must understand the characteristics of the score. What are the types of errors that the score is most sensitive to? Does the score encourage biased forecasts? Are false alarms punished more (or less) than missed events? Does the score behave in the same way for rare events as it does for more common events? Sensitivities of this type for accuracy measures have been considered by several reserachers over the past several decades.

*Corresponding author address: Michael E. Baldwin, CIMMS/OU, 1313 Halley Cir, Norman, OK, 73069
Email: mbaldwin@ou.edu

Mason (1989) examined the sensitivity of the threat score (critical success index) to the observed event frequency as well as the decision threshold, which can be related to the frequency bias. His results showed that the threat score is highly sensitive to both factors. Common events result in higher threat scores than rare events, and the threat score is maximized for bias values greater than 1 (overforecasting). Hamill (1999) discussed the implications of these sensitivities in determining confidence intervals (error bars) for the threat score. Other researchers have proposed modifications to the threat score to attempt to reduce the sensitivity to forecast bias and event frequency (Schaefer 1990; Mesinger and Brill 2004).

Other researchers have focused on the "fairness" of scores. For example, an accuracy measure is defined as *equitable* by Gandin and Murphy (1990) if the same score (usually zero) is given to either a random or constant forecast. Equitable scores do not encourage over- or under-forecasting of an event, therefore one assumes that the score will be maximized when the frequency bias is equal to 1. However, Marzban (1998) could not find an equitable score for rare events under realistic forecast conditions. Only a few scores are maximized when bias = 1 under very specialized conditions; when the variance of forecast values associated with "no" events is equal to that associated with "yes" events.

A broader question is: for what forecast situations *should* scores be maximized for bias = 1? This will depend on how the forecasts are used and how different outcomes affect the user of the forecast. This relates to the issue of forecast *value*. Forecast value is defined by Murphy (1993) as the benefits of forecast information to a user of the forecast. Each user will have a different level of sensitivity to false alarms and missed events, depending on their individual situation. For certain situations, a biased forecast may, in fact, be more valuable than an unbiased foreacst. Thornes and Stephenson (2001) provide an example of the complicated relationship between forecast bias, accuracy, and value for a winter weather forecasting

situation. The cost/loss situation for a city deciding whether or not to treat slippery roads was analyzed for two competing forecast providers. Thornes and Stephenson (2001) found that a forecast provider with a bias of 1.8 resulted in greater value to the city than one with a bias of 1, even though other accuracy measures (percent correct, false alarm rate) showed the unbaised forecast to be preferred. One might question whether forecasters should be concerned with forecast value, since they have no control over the decisions made by the users of forecast information. Since typical forecasts (of hazardous weather in particular) provide only yes/no information, it is up to those that issue such forecasts to consider their value for the variety of users. End-users of forecast information would likely find verification information related to forecast value to be quite useful.

Ideally, one might desire to decompose forecast errors into separate independent factors. For example, Murphy (1996) shows how scores related to the mean square error can be decomponsed into components due to bias, reliability, and resolution. For spatial forecasts in particular, one could consider several components of forecast error. Ebert and McBride (2000) describe a technique to decompose errors in precipitation forecasts into components due to displacement, amplitude (bias), and shape errors. In this paper, we address spatial forecast errors such as those associated with numerical forecast precipitation. The sensitivity of several measures of accuracy to bias and displacement errors will be examined for a hypothetical forecasting situation.

## 2. HYPOTHETICAL FORECAST SITUATION

The spatial forecast situation, such as one might face when forecasting precipitation greater than a given threshold, will be modelled using a simple hypothetical example. Here, regions of forecast and observed "yes" fields will be represented by circular shapes. The observed circle will have a radius = $r_o$ (fixed $r_o = 1$), the forecast circle will have a radius = $r_f$ and the circles will be displaced by a distance = **D** (Fig. 1). Since the area of the observed circle is fixed, the bias error will be varied by varying the radius of the forecast circle. As can be seen in figure 1, the frequency bias (**B** = frequency of forecast = yes divided by frequency of observed = yes) simplifies to $r_f * r_f$ (since $r_o = 1$). For a fixed forecast circle (or fixed **B**), as **D** increases, the area of overlap will decrease until it reaches zero at **D** = $r_o + r_f$. For a fixed displacement **D**, as the bias increases, the overlap area will increase until the forecast circle envelops the observed circle completely. The observed event frequency is the ratio of the observed area to the total forecast domain. The bias cannot be larger than the inverse

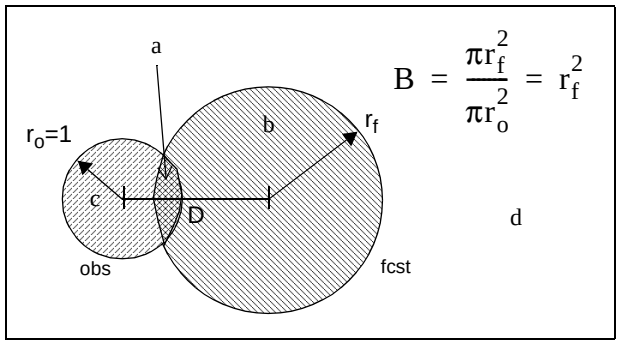


Figure 1: A hypothetical example of a spatial forecast and associated observation. The radius of the observed circle ($\mathbf{r_o}$) is fixed at 1, the radius of the forecast circle is $\mathbf{r_f}$, the distance between the centers of the observed and forecast circles is **D**. The area of overlap between the observed and forecast circles is **a** (forecast = yes & observed = yes). The frequency bias, or ratio of the forecast and observed areas, is **B**. The total forecast domain is indicated by the outer rectangle. The elements of the contingency table are indicated by the corresponding area on the figure (**b**: forecast = yes & observed = no, **c**: forecast = no & observed = yes, **d**: forecast = no & observed = no).

of the event frequency.

Since this is a binary (yes/no) type of forecast, it can be verified through the use of a 2x2 contingency table. The scores that will be analyzed in this paper will be defined through the use of this table:

**Table 1: Contingency table for a given event.**

| | | Observed | | |
|---|---|---|---|---|
| | | yes | no | Total |
| Forecast | yes | a | b | a+b |
| | no | c | d | c+d |
| | Total | a+c | b+d | N |

In this paper, we will show how a variety of scores vary as a function of bias and displacement errors. Therefore, the various elements of the contingency table must be computed as a function of **B** and **D**. Given the forecast and observed circle areas (defined by their radii), the area of the intersection, and the frequency of the observed event, each element of the contingency table can be computed analytically. The intersection of the forecast and observed circle represents the "hit" area **a**:

$$a = r_o^2 \cdot \operatorname{acos}\left(\frac{D^2 + r_o^2 - r_f^2}{2Dr_o}\right) + r_f^2 \cdot \operatorname{acos}\left(\frac{D^2 + r_f^2 - r_o^2}{2Dr_f}\right)$$
$$-\frac{1}{2}\sqrt{(r_f + r_o - D)(D + r_f - r_o)(D + r_o - r_f)(D + r_f + r_o)}$$

Since the radius of the observed circle is fixed at 1.0, using the definition of the frequency bias, this equation can be simplified and written to be a function of

bias and displacement errors:

$$a = \operatorname{acos}\left(\frac{D^2 + 1 - B}{2D}\right) + B \cdot \operatorname{acos}\left(\frac{D^2 + B - 1}{2D\sqrt{B}}\right)$$

$$-\frac{1}{2}\sqrt{(\sqrt{B} + 1 - D)(D + \sqrt{B} - 1)(D + 1 - \sqrt{B})(D + \sqrt{B} + 1)}$$

The area of the observed "yes" circle is:

$$a + c = \pi r_o^2 = \pi$$

and the area of the forecast "yes" circle is:

$$a + b = \pi r_f^2 = \pi B$$

Therefore, the **b** and **c** contingency table elements can be written in terms of **D** and **B**:

$$b = \pi B - a$$

$$c = \pi - a$$

where **a** is defined above. Finally, the **d** element (forecast = no and observed = no) depends upon the size of the total forecast domain, which in turn controls the frequency of the event, **p**. The sensitivity of scores to this factor will also be examined.

$$p = \frac{a + c}{N} = \frac{\pi}{N}$$

$$N = \frac{a + c}{p} = \frac{\pi}{p}$$

$$d = N - a - b - c = a + \pi\left(\frac{1}{p} - B - 1\right)$$

### 3. DEFINITION OF ACCURACY MEASURES

Several commonly used measures of forecast accuracy will be examined. The elements of the 2x2 contingency table have been computed in terms of bias and displacement errors for the hypothetical situation described above. The definitions of these scores are as follows:

$$\text{Probability of detection} = POD = \frac{a}{a + c}$$

$$\text{Threat score} = TS = \frac{a}{a + b + c}$$

$$\text{Equitable threat score} = ETS = \frac{a - a_{rand}}{a + b + c - a_{rand}}$$

$$a_{rand} = \frac{(a + b)(a + c)}{N}$$

$$\text{True Skill Statistic} = TSS = \frac{a}{a + c} - \frac{b}{b + d}$$

$$\text{Heidke Skill Score} = HSS = \frac{a + d - E}{N - E}$$

$$E = \frac{(a + b)(a + c) + (b + d)(c + d)}{N}$$

$$\text{Odds ratio skill score} = ODDS = \frac{\theta - 1}{\theta + 1}$$

$$\text{Odds ratio} = \theta = \frac{ad}{bc}$$

Since these scores are fairly common, only a brief description and associated references will be provided here. POD is simply the fraction of the observed events that were correctly predicted. TS is the fraction of the observed and predicted events that were correct (Gilbert 1884 called this the "ratio of verification"). ETS adjusts the threat score to remove the number of correct yes forecasts expected due to random chance (Schaefer 1990 called this the "Gilbert skill score"). TSS is the probability of detection minus the probability of false detection (Doswell et al 1990 called this the True Skill Statistic, while Peirce 1884 may have been the first to discover it; it has also been called the Kuipers performance index and the Hanssen-Kuipers discriminant). HSS is similar to the ETS, except it is the fraction of both the correct yes and no forecasts divided by the total forecast domain, with the correct forecasts adjusted to remove those expected by random chance. ODDS is a function of the odds ratio and varies from -1 to 1, a random forecast results in ODDS = 0 (Stephenson 2000).

Two other scores are also included which are related to forecast value. The value of a forecast depends upon user requirements. Thompson and Brier (1955) proposed the use of a simple cost/loss ratio for estimating value. The first score is a modified version of the threat score by Donaldson et al. (1975), who introduced a *k* factor to the TS formula, which the false alarm element in the contingency table was divided by. This factor was meant to represent the loss/cost ratio, since the cost/loss ratio is typically presented in analysis of value, we have modified the definition to use the cost/loss (C/L) ratio directly:

$$\text{Critical success index with k} = CSIK = \frac{a}{a + \frac{C}{L}b + c}$$

The second score, called the value index (Thornes and Stephenson 2001) is a simplified way of estimating the relative value of a forecast. In the 2x2 contingency table, the expenses due to protective action (cost=C) and the expenses resulting from damage due to the weather event and no protective action (loss=L) are given in Table 2.

One assumes that protective action is taken whenever the forecast is "yes" for the observed event. The cost of protective action is = C. If no protective action is taken and the observed event occurs, the resulting loss = L. If no event occurs and no protective

**Table 2: Expenses associated with forecast information.**

| | | Event occurred? | |
|---|---|---|---|
| | | yes | no |
| Action taken? | yes | C | C |
| | no | L | 0 |

action is taken, there is no expense. Therefore, the total expense resulting from the use of a forecast with errors found in the contingency table are:

$$E(\text{forecast}) = aC + bC + cL$$

The relative value of a forecast is the benefit (expense savings) of using a forecast divided by the benefit provide by a perfect forecast:

$$\text{Value Index} = \text{VALUE} = \frac{E(\text{no forecast}) - E(\text{forecast})}{E(\text{no forecast}) - E(\text{perfect})}$$

For a perfect forecast, all observed events are correctly forecast, therefore **a**=p. The expenses associated with a perfect forecast are:

$$E(\text{perfect}) = pC$$

In this work, we assume that a user with no forecast information takes action depending upon the climatological frequency of the event. If the event occurs frequently relative to the cost/loss ratio, the user will always take action to protect. No losses will occur, but the expenses will be due to the cost of protection (N*C). If the event is rare relative to the cost/loss ratio, the expenses will be less if the user never takes protective action. Expenses will occur as losses for each observed event. Therefore, the expenses associated with no forecast information depend on p and C/L:

$$E(\text{no forecast}) = pL \text{ if } p \le \frac{C}{L}$$

$$E(\text{no forecast}) = NC \text{ if } p > \frac{C}{L}$$

Using these expenses, the value index functions become:

$$\text{VALUE} = \frac{\frac{a}{C/L} - (a+b)}{(a+c)\left(\frac{1}{C/L} - 1\right)} \text{ if } p \le \frac{C}{L}$$

$$\text{VALUE} = \frac{c + d - \frac{c}{C/L}}{b+d} \text{ if } p > \frac{C}{L}$$

## 4. RESULTS

Figures 2 and 3 show how various accuracy measures vary as a function of bias and displacement errors for a somewhat rare observed event (p=0.05). This event frequency is similar to fractional area of preciptitation observed greater than 0.5" analyzed to a 40km x 40km grid box across the contiguous 48 states during the warm season. For a fixed displacement error, the POD increases as bias increases, which is expected since the forecast circle enlarges until it eventually "swallows" the observed circle and POD=1. For a fixed bias error, the POD decreases as displacement error increases, since the circles are moving away from each other the overlap area will eventually decrease to zero. TS and ETS are practically equivalent for rare events, and have similar behavior as HSS. The maximum score is found for **B**>1 for all **D**>0, and the **B** that is associated with the maximum score increases as **D** increases. TSS has a similar behavior, except the maximum score axis increases linearly in **B** as **D** increases. ODDS cannot be computed if the **b** or **c** elements are zero, such as the case when POD=1 or there are no false alarms (the situation where the forecast circle is completely contained within the observed circle, for small values of **B** and **D**). ODDS appears to be less sensitve to changes in **B** than the other scores.

Figures 4 and 5 show the various accuracy measures as a function of bias and displacement errors for a somewhat more common event (p=0.33). This event frequency is similar to that observed for precipitation greater than 0.01" analyzed to a 40km x 40km grid box. As expected, for a fixed displacement error, the POD increases as bias increases. TS and ETS are considerably different for common events. The maximum score axis for TS slopes in the positive **B** direction, as it did for the rare event. However, the maximum score for ETS is fairly constant and near **B**=1. For a fixed **D**, ETS appears to be less sensitive to changes in **B** for common events. HSS and TSS behave similarly to ETS. Again, ODDS is quite different from the other scores, with little if any variation in score for fixed **D** as **B** changes. For most of these scores in the common event case, the scores appear to be fairly insensitive to bias and produce maxima near **B**=1.

Figure 6 shows the value-related measures as a function of bias and displacement errors for a relatively low cost/loss ratio = 0.1. The top panels are for a rare event (p=0.05) and the bottom panels are for a common event (p=0.33). For the low cost/lost ratio situation, CSIK and VALUE provide consistent information, with a maximum value axis sloping in the positive **B** direction. This indicates that for forecasts with any displacement error, a frequency bias greater than 1 will provide a more valuable forecast. The accuracy measures in figure 2 provide information consistent with the value for the rare event situation. However, except for the TS, the accuracy measures in the common event situation show

that the maximum *accuracy* is associated with **B**=1, while the maximum *value* is obtained for forecasts with **B**>1. These plots indicate that users in the low cost/loss ratio situation are very sensitive to missed events. The cost of taking action is relatively low, therefore false alarms are not punished as much as missed events. Therefore, a forecast with a large bias is considered valuable in this situation.

Figure 7 shows the value-related measures as a function of bias and displacement errors for a relatively high cost/loss ratio = 0.5. Again, the top panels are for a rare event (p=0.05) and the bottom panels are for a common event (p=0.33). For the high cost/lost ratio situation, CSIK and VALUE do not provide consistent information. CSIK is very consistent with the TS plots in figures 2 and 3, with a maximum score axis sloping in the positive **B** direction. However, the VALUE plots indicate the opposite behavior, as **D** increases, the maximum value is provided for smaller and smaller bias errors. This indicates that for forecasts with any displacement error, a frequency bias less than 1 will provide a more valuable forecast. The high cost/loss ratio situation is associated with users that are very sensitive to false alarms, taking action when no event is observed is very costly to this type of user. To reduce the number of false alarms, a forecast with low bias must be produced. This is true in both the rare event and in the common event situations. None of the accuracy measures in figures 2 and 3 provide information consistent with the value for the high cost/lost ratio situation.

## 5. SUMMARY

No single score can provide perfect information on forecast quality to satisfy all types of users. The sensitivity of various measures of forecast accuracy and value is analyzed for a hypothetical forecast situation. Most scores are found to be quite sensitive to bias error, event frequency, and displacement error. The odds ratio skill score is the least sensitive to bias error and event frequency. For rare events with any displacement error, accuracy measures are maximized for bias greater than 1. For common events, most of the scores are maximized at bias = 1 over a wide range of displacement errors.

Most accuracy measures, as well as the modified CSI of Donaldson et al. (1975) provide information that is consistent with value for rare events and low cost/loss ratios. For common events, the TS and modified CSIK provide information consistent with value. However, in the high cost/loss ratio situation, none of the accuracy measures provide information consistent with the estimate of value.

For low cost/loss ratio users, missed events are critical, therefore a forecast with a high bias provides more value than an unbiased forecast. The opposite is true for high cost/loss users, false alarm errors are the most critical, therefore a forecast with low bias provides more value than an unbiased forecast.

## References


Brooks, H.E. and C.A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288-303.

Donaldson, R. J., R. M. Dyer, and M. J. Krauss, 1975: An objective evaluator of techniques for predicting severe weather events. *Preprints, 9th Conf. Severe Local Storms*, Norman, OK, Amer. Meteor. Soc., 321-326.

Doswell, C.A. III, R. Davies-Jones, and D.L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576-585.

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

Gandin, L. S. and A. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370.

Gilbert, G. F., 1884: Finley's Tornado Predictions. *American Meteorological Journal*, **1**, 166-172.

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

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753-763.

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

Mesinger, F. and K. Brill, 2004: Bias normalized precipitation scores. *Preprints*, *17th Conf. on Probability and Statistics*, Amer. Meteor. Soc., Seattle, WA, paper J12.6.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science,* **4,** 453–454.

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.

Thompson, J. C. and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83**, 249-254.

Thornes, J.E. and D.B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteorol. Appl.*, **8**, 307-314.


.

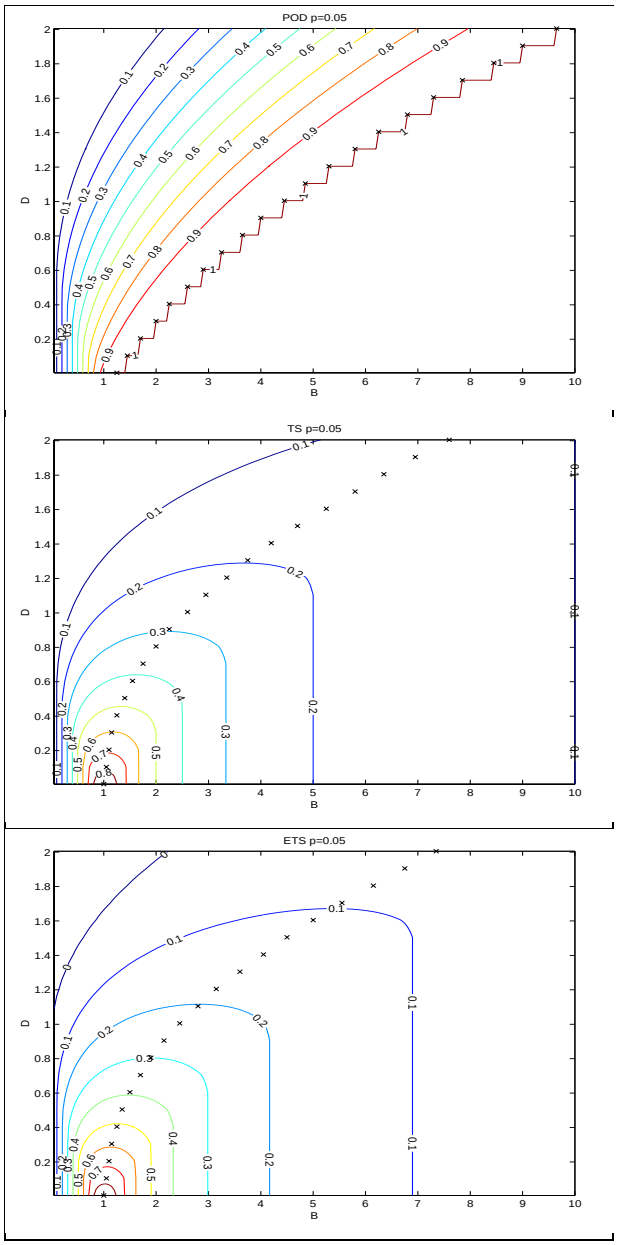


Figure 2: For event frequency p=0.05, accuracy measures as a function of bias error (B) and displacement error (D) for the circle-circle interaction situation. Top panel is POD, middle is TS, bottom is ETS. Axis of maximum score value is indicated by x's.

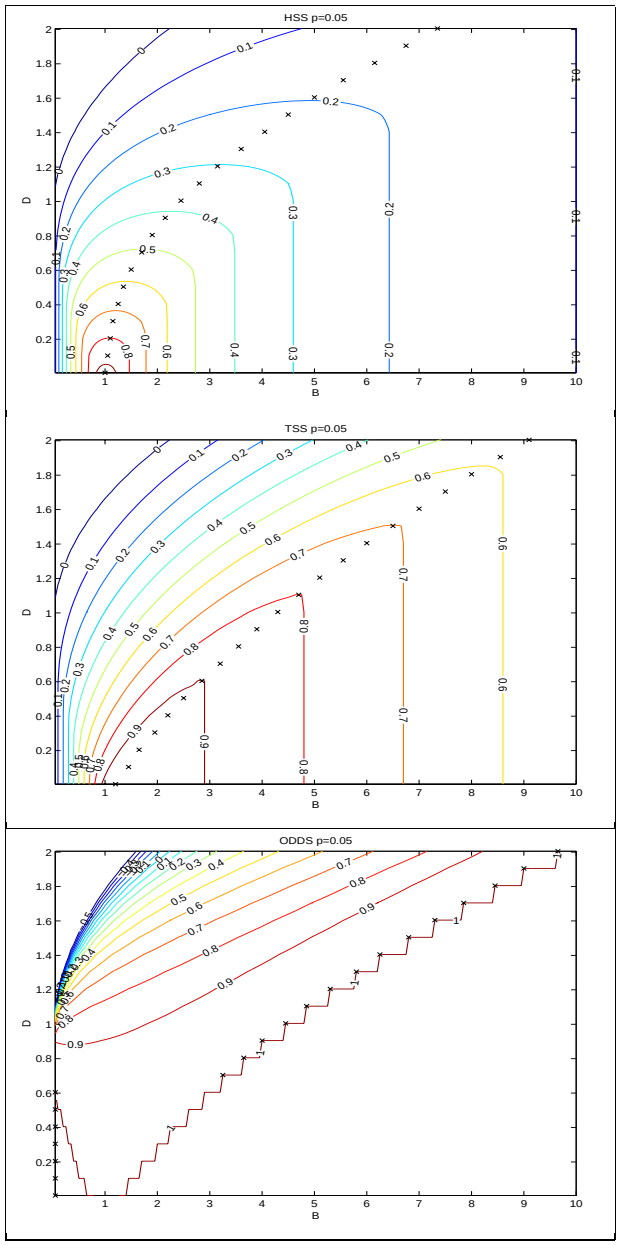


Figure 3: As in figure 2, except top panel is HSS, middle panel is TSS, and bottom panel is ODDS.

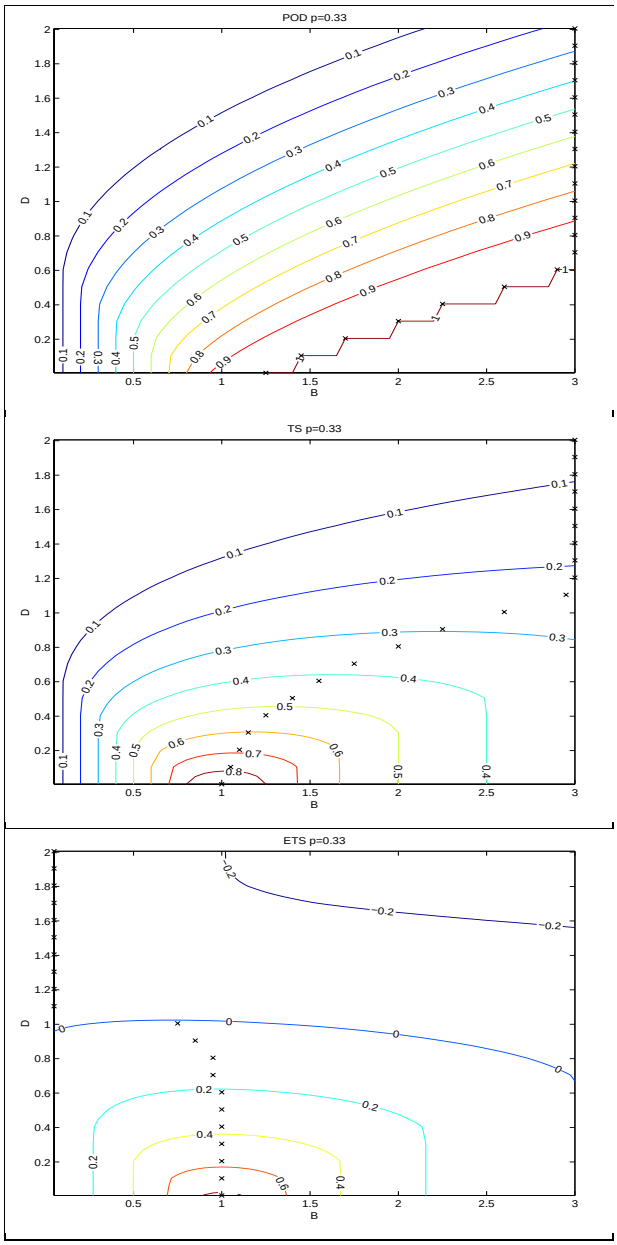


Figure 4: As in figure 2, except for p=0.33.

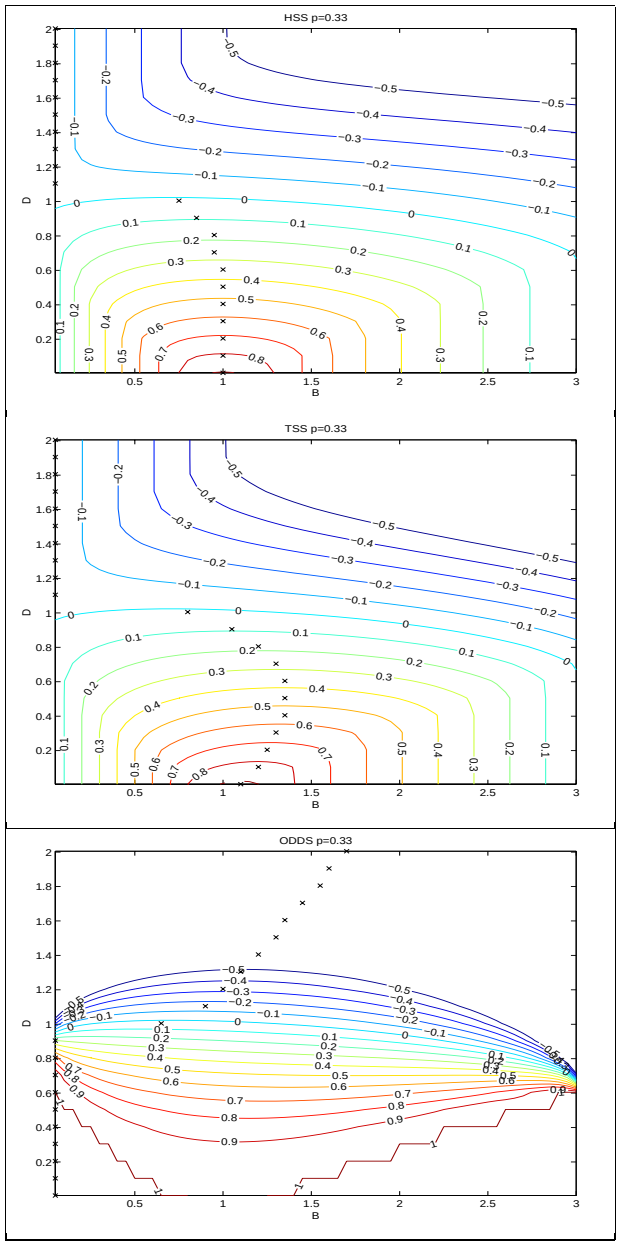


Figure 5: As in figure 3, except for p=0.33.

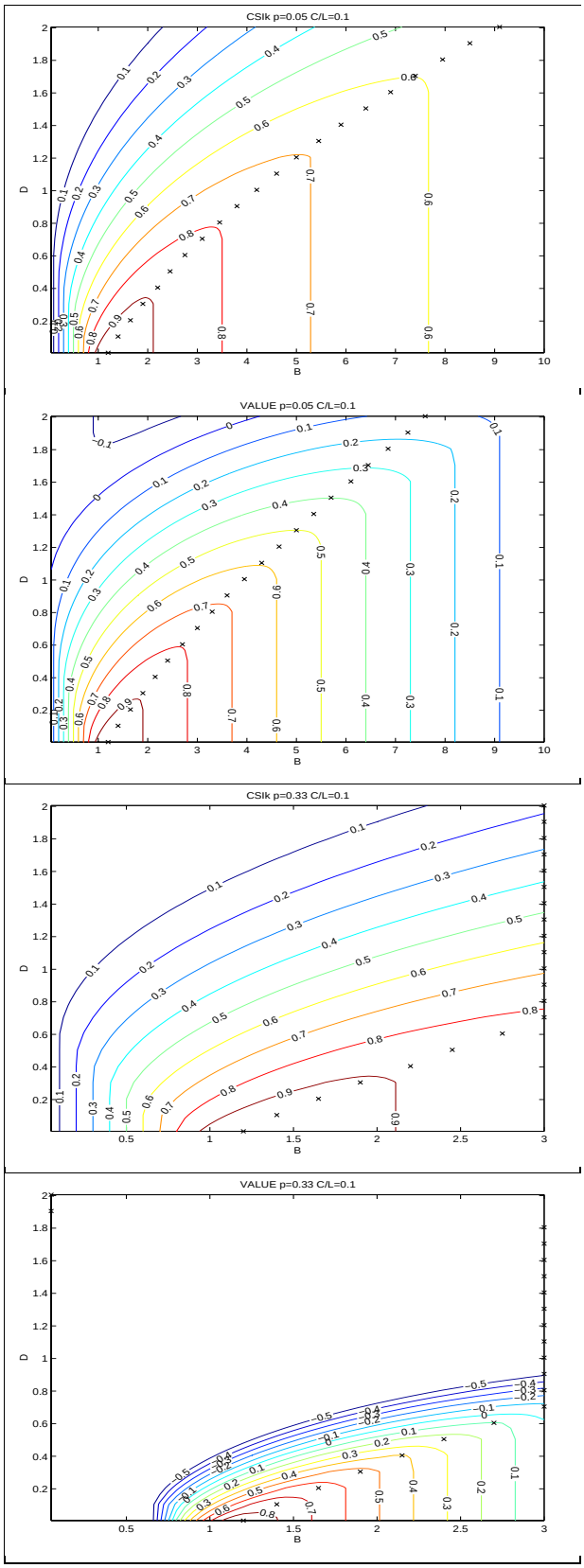


Figure 6: Value-related scores for C/L=0.1 as a function of bias and displacement errors. Top panel is CSIK, second panel is VALUE for p=0.05, third panel is CSIK, bottom panel is VALUE for p=0.33.

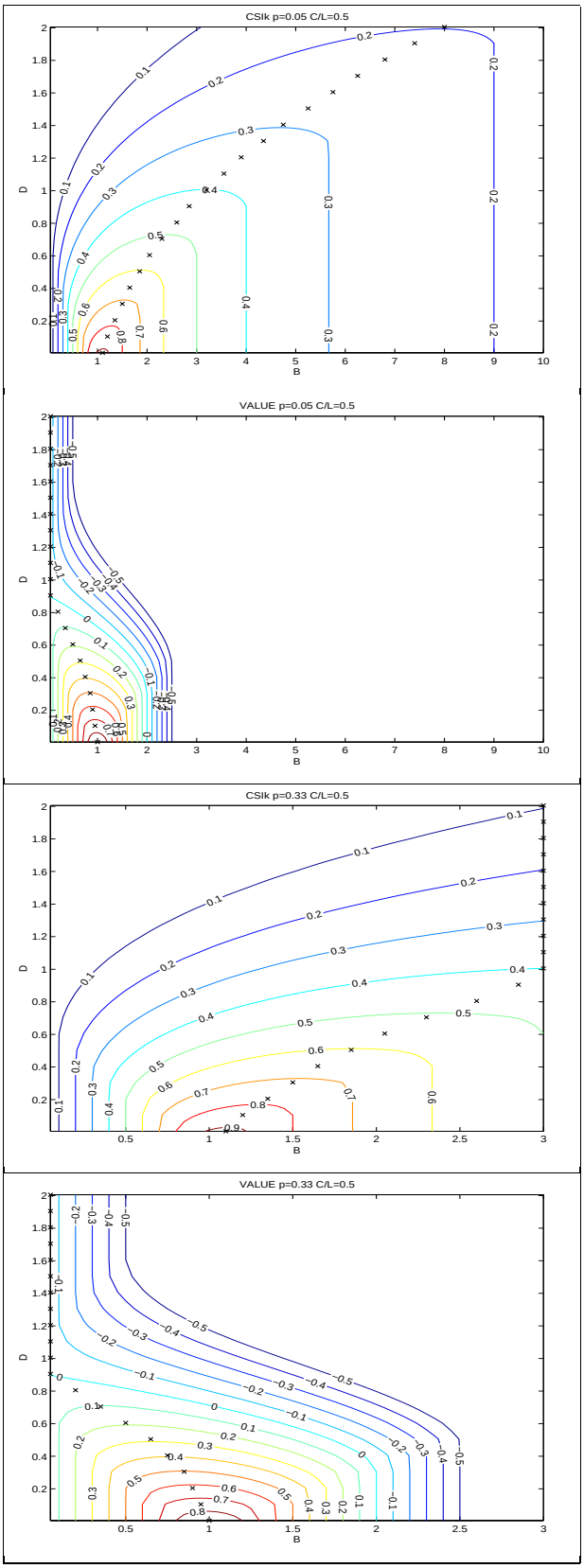


Figure 7: As in figure 6 except for C/L=0.5.