The NWRA Flare-Forecast Comparison Workshops: Approaches for Meaningful Verification and Comparison

G. Barnes, K.D. Leka and the International Flare Forecasting Comparison Group

Motivation for the Workshops

- Leka & Barnes (2007) “conclude that the state of the photospheric magnetic field at any given time has limited bearing on whether that region will be flare productive.”
- Georgoulis & Rust (2007) “find that $B_{\rm eff}$ is a robust criterion for distinguishing flaring from nonflaring regions.”
- McAteer et al. (2005) state “the fractal dimension does not fully capture the relationship between active region complexity and flare rate.”
- Schrijver (2007) states “Clearly, R is a far more significant measure for major flare potential than the unsigned flux”.
- Who is correct, and perhaps equally importantly, what is meant by a clear or robust measure for distinguishing flaring from nonflaring active regions?

Overview

What should be accounted for in making meaningful comparisons of flare forecasting methods?

- Use of skill scores versus accuracy.
- Estimates of the random error.
- Definition of event (magnitude, validity, latency).
- The set of days/times/active regions for which predictions are made.
- What is a method actually trying to optimize for?

Skill Scores versus Accuracy

Low event rates (typical of large solar events) mean that it's easy to get a high accuracy by forecasting that nothing ever happens.
E.g., 2009 workshop: M5.0 and above, 12 hr window, event rate=0.007, so accuracy of 0.993 is easy. Instead, consider the skill.

- Skill: relative performance with respect to a reference forecast
- Skill score calculation:
  - Then the skill score, SS, is:

$$SS = \frac{M - M_{\mathrm{ref}}}{M_{\mathrm{perfect}} - M_{\mathrm{ref}}}$$

- If the forecasts are perfect, $SS = 1$
- If the forecasts are no better than the reference, $SS = 0$
- If the forecasts are worse than the reference, $SS < 0$

Skill Scores versus Accuracy

How big an effect is this? Some examples:

	Accuracy	Heidke Skill Score
Climatology	0.993	0.000
Method 1	0.984	0.190
Method 2	0.993	0.011
Method 3	0.993	1.1×10^{-5}

- Note that the method with the lowest accuracy has the highest Heidke skill score. (This is not a general property.)
- This largely explains the differences in the conclusions about the success of flare forecasting methods.
- To 3 significant figures, none of the methods improve on the accuracy of forecasting no event.
    - Is this a significant difference?

Estimates of the Random Error

Suppose we have a set of n observations: x_1, x_2,. . . ,x_n, and want to make a prediction about a new observation. Estimate the uncertainty in the resulting predictions using a bootstrap method (jackknife is a similar alternative (see Efron & Gong, 1983, The American Statistician, 37, 36–48, “A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation.”):

- Draw at random with replacement from the set of observations a new set of n (probably not distinct) observations x^*_1, x^*_2,. . . ,x^*_n.
- Compute whatever quantity one is interested in (e.g., a skill score) from this new set.
- Repeat this process a large number of times, and use the resulting distribution of the quantity to estimate the most likely value of and uncertainty in the quantity.

For the previous example, a typical uncertainty in the accuracy is less than 0.001, while in the skill score it is 0.03.

Definition of Event

For workshops, we used several different definitions of event. For example

- C1.0 and above, 24 hr
- M1.0 and above, 12 hr
- M5.0 and above, 12 hr

How do the results change with event definition?

	C1.0	M1.0	M5.0
Method 1	0.43	0.34	0.19
Method 2	0.47	0.08	0.01
Method 3	0.42	-2×10^{-6}	1×10^{-5}
Uncertainty	0.01	0.03	0.03

Even for Method 1, which had the most consistent performance (skill scores), the differences between event definitions are likely significant. To make meaningful comparisons, the same event definition should always be used.

Prediction Set

It is common for a method to restrict forecasts to only certain kinds of input data or time intervals. For example, one method considered in the workshops only produces a forecast within 30° of solar disk center.

For that subset of data:

HK/P/T SS:	0.21
Brier SS:	0.19

When “reference forecast” is used to include all data in standard set:

HK/P/T SS:	0.07
Brier SS:	0.06

Similar issues may occur when different time intervals are considered, for example because of solar cycle variations, or when different active region extraction algorithms are used, for example HMI Active Region Patches versus NOAA Active Regions.

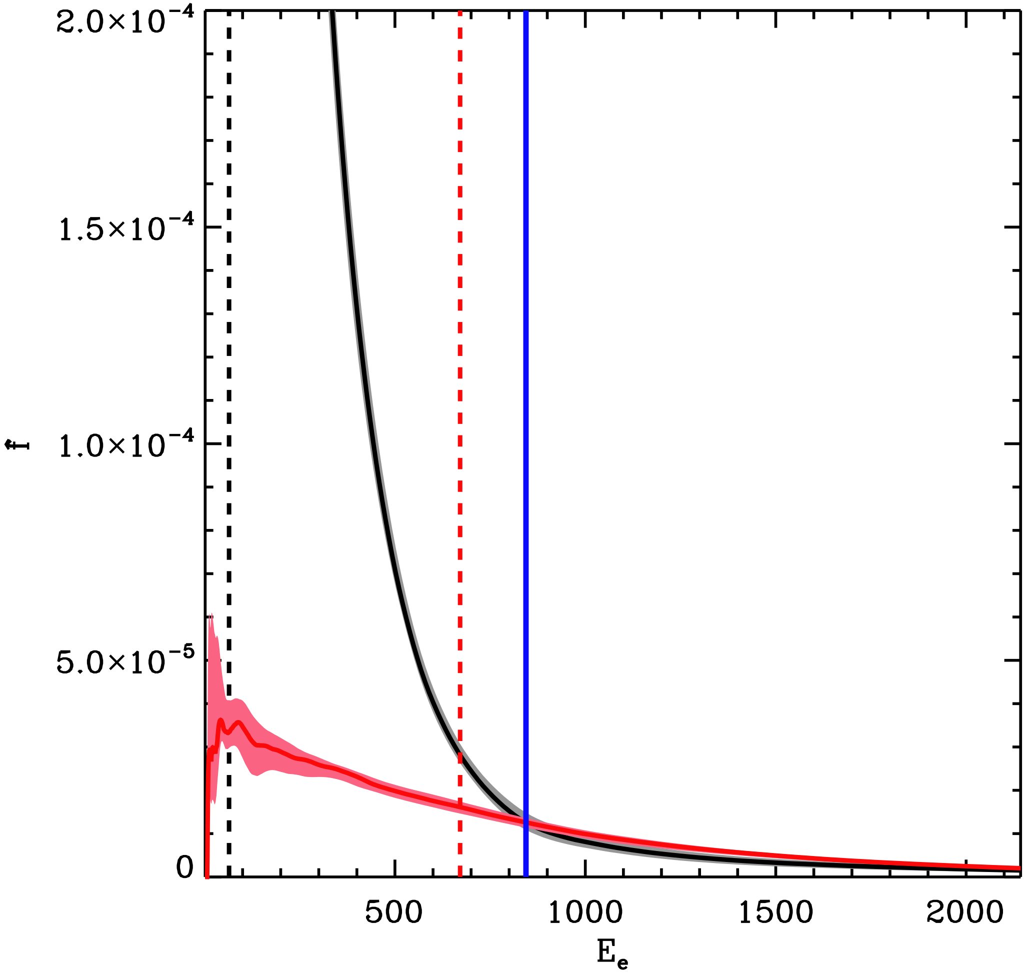
What are Methods Trying to Optimize?

Discriminant Analysis (typically) maximizes the accuracy:

	observed	
predicted	event	no event
event	27	66
no event	19	3227

Accuracy=0.976 ± 0.003

Peirce Skill Score=0.32 ± 0.07



What are Methods Trying to Optimize?

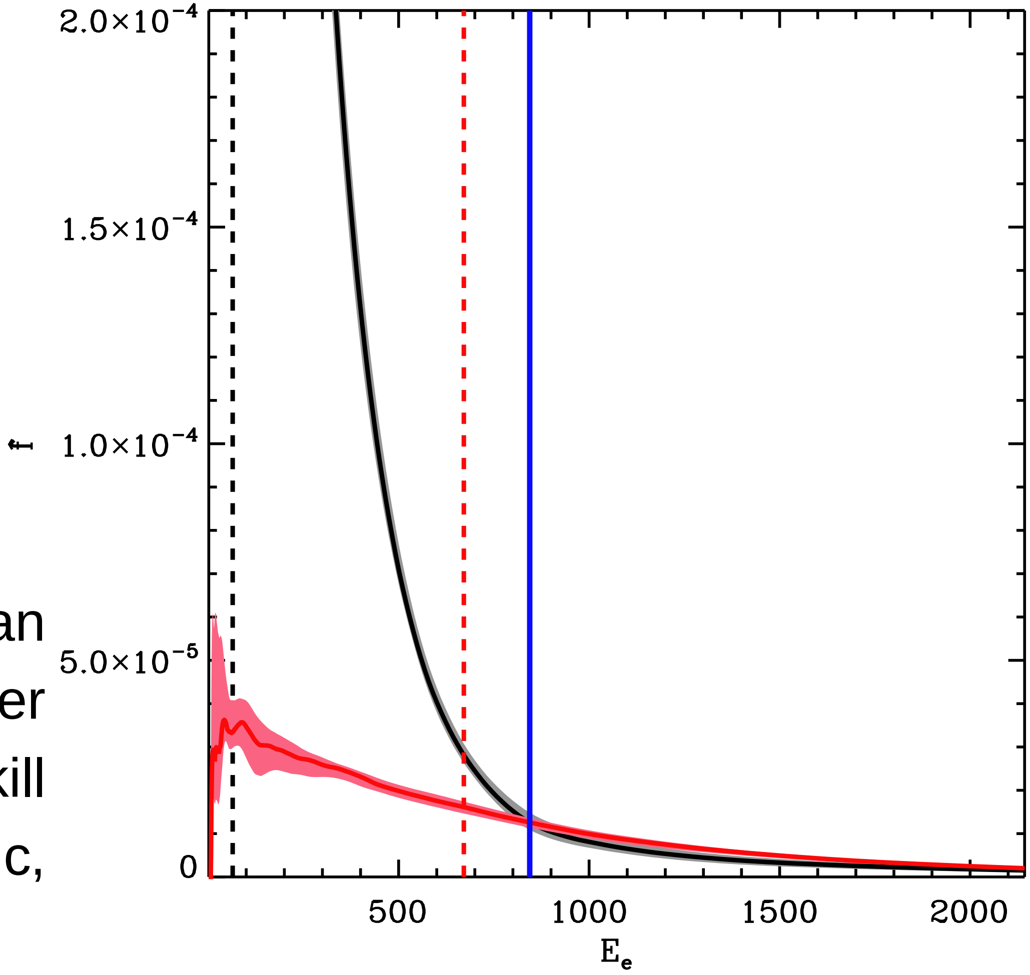
Discriminant Analysis (typically) maximizes the accuracy:

	observed	
predicted	event	no event
event	27	66
no event	19	3227

Accuracy=0.976 ± 0.003

Peirce Skill Score=0.32 ± 0.07

However, discriminant analysis can be adjusted to maximize other things, such as the Peirce Skill Score (a.k.a. True Skill Statistic, Hanssen & Kuipers discriminant):



PSS = probability of detection - false alarm rate

What are Methods Trying to Optimize?

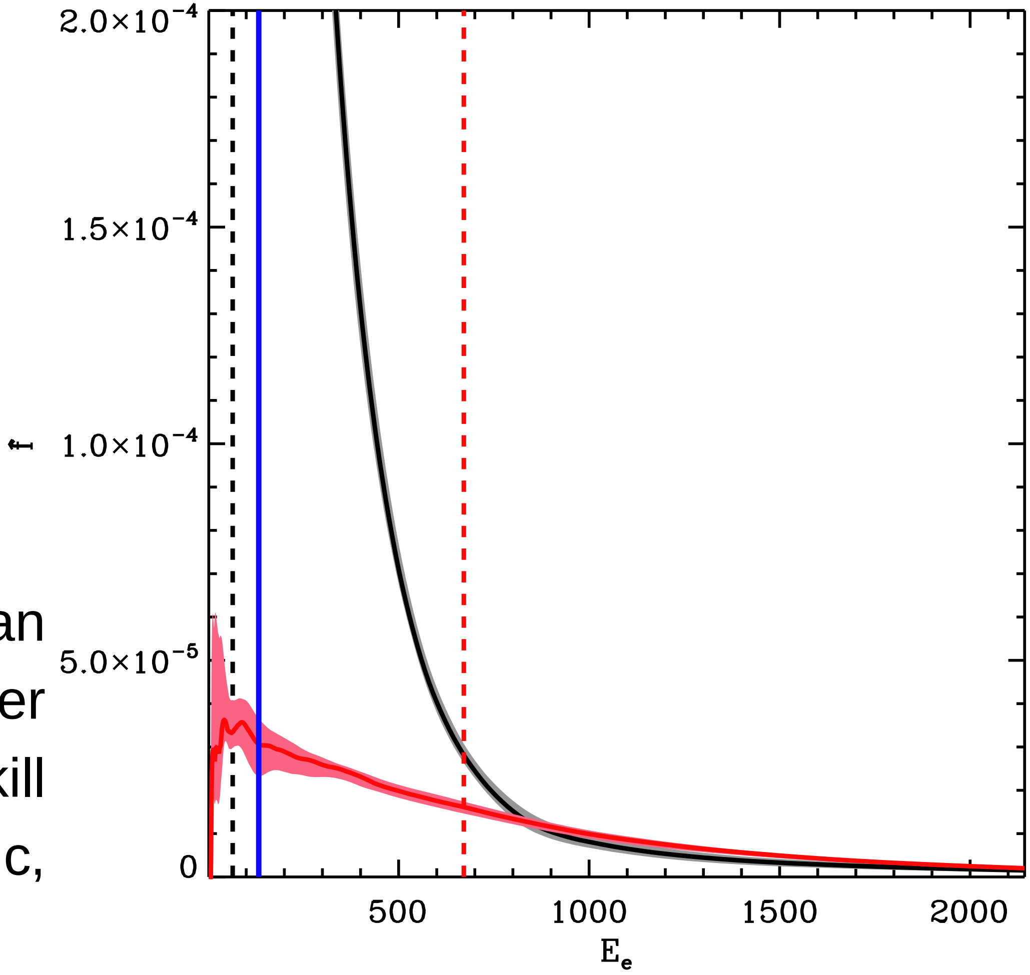
Discriminant Analysis (typically) maximizes the accuracy:

	observed	
predicted	event	no event
event	27	66
no event	19	3227

Accuracy=0.976 ± 0.003

Peirce Skill Score=0.32 ± 0.07

However, discriminant analysis can be adjusted to maximize other things, such as the Peirce Skill Score (a.k.a. True Skill Statistic, Hanssen & Kuipers discriminant):



PSS = probability of detection - false alarm rate

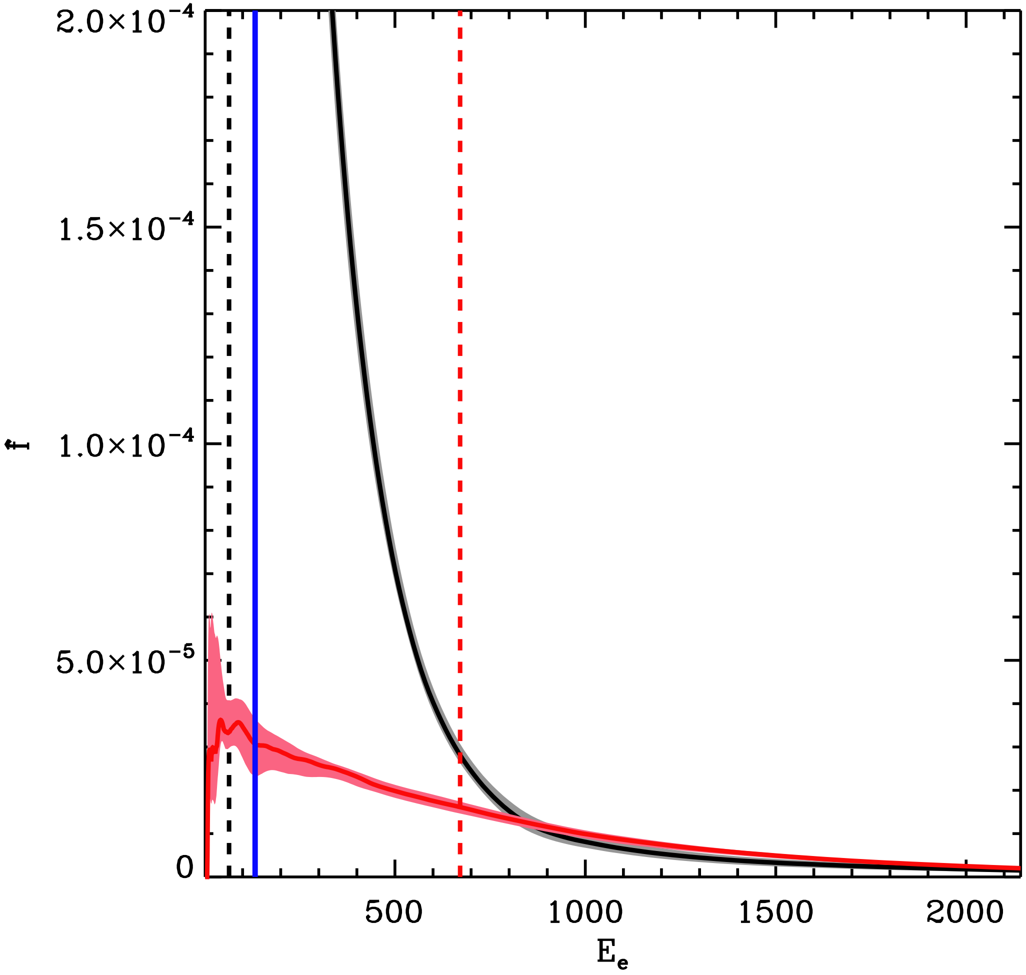
What are Methods Trying to Optimize?

Discriminant Analysis (typically) maximizes the accuracy:

	observed	
predicted	event	no event
event	27	66
no event	19	3227

Accuracy=0.976 ± 0.003

Peirce Skill Score=0.32 ± 0.07



	observed	
predicted	event	no event
event	79	14
no event	417	2829

Accuracy=0.905 ± 0.008 Peirce Skill Score=0.72 ± 0.04

The result is a much lower accuracy, but a much higher Peirce Skill Score.

Conclusions

- Estimates of the random error are important.
- Consistency is paramount.
    - The same definition of event (magnitude, validity, latency) should be used.
    - Predictions should be made for the same set of days/times/active regions.
    - Ensure methods are actually trying to optimize for the same quantity.
- For more on the workshops, see Poster 10.07.