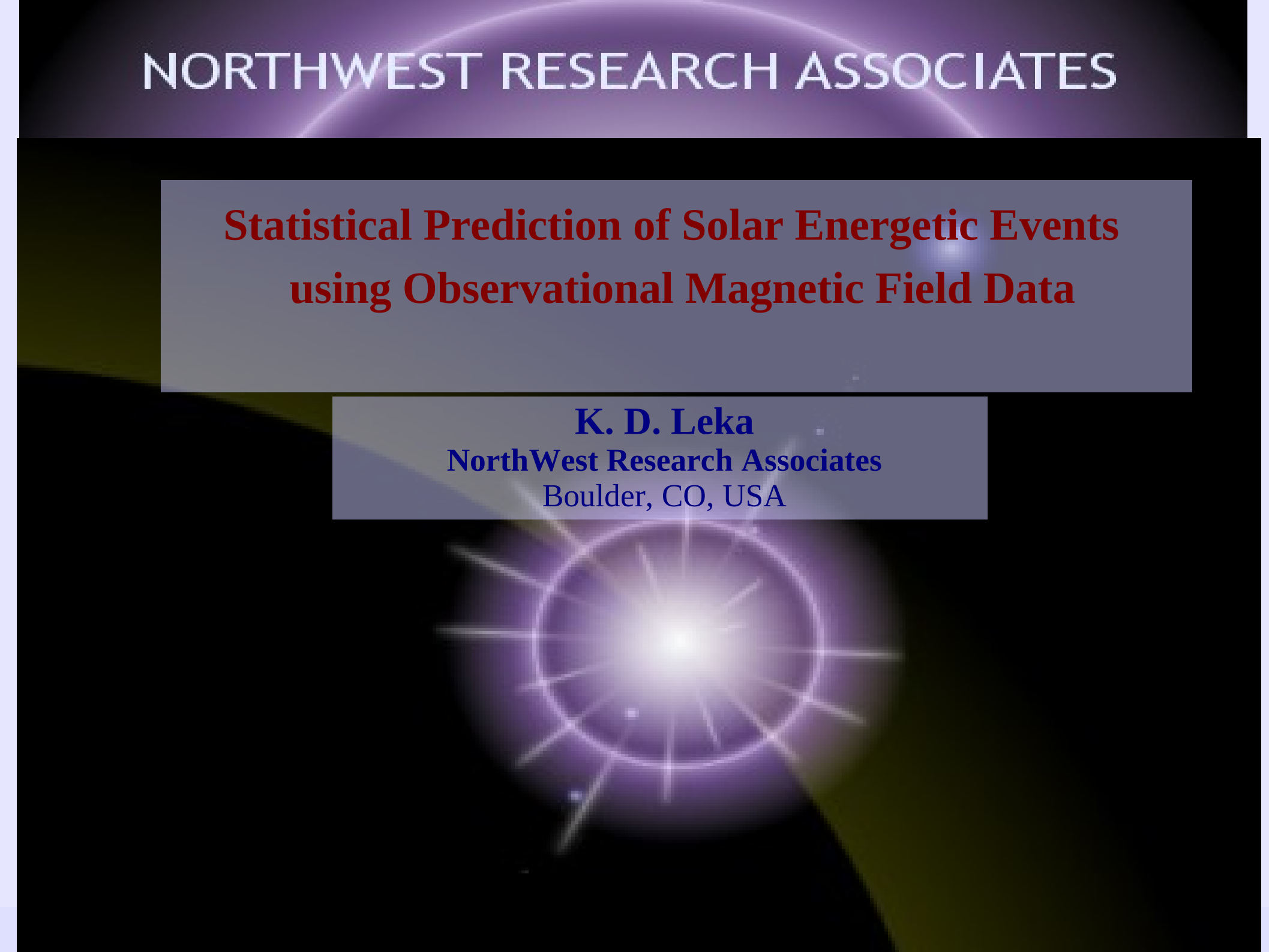
NORTHWEST RESEARCH ASSOCIATES

Statistical Prediction of Solar Energetic Events using Observational Magnetic Field Data

K. D. Leka
NorthWest Research Associates
Boulder, CO, USA



Introduction: What, who cares and why

Data source(s) and analysis techniques

An introduction to Discriminant Analysis

But first....

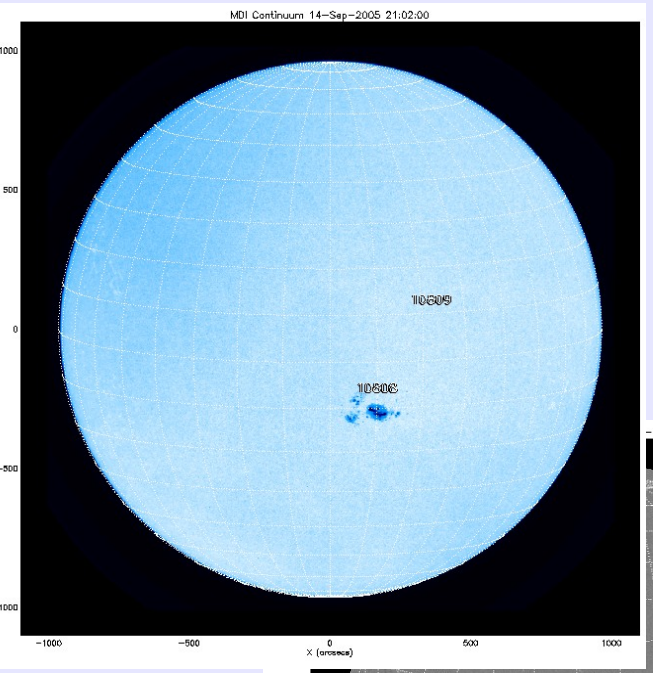
Collaborators, Colleagues, and Funding Agencies

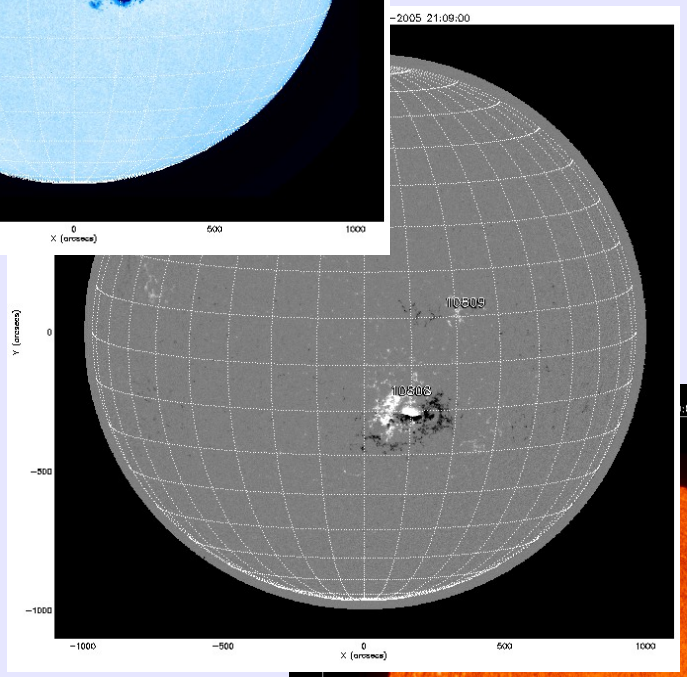
Graham Barnes
Richard Canfield
Devin Della-Rose
Yuhong Fan
Steve Gurteslough
Gary Heckman
Barry LaBonte
Dana Longcope
Tom Metcalf
Don Mickey
Bill Murtaugh
Gary Nitta
Evelyn Schumer
Mark Waterson
The forecasters at NOAA/SEC

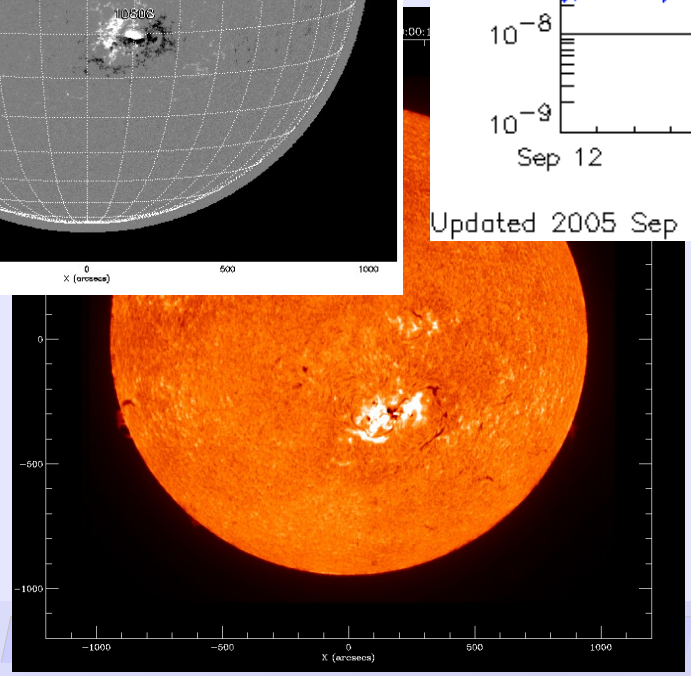
Air Force Office of Scientific Research
National Research Council
NASA
NOAA/Space Environment Center

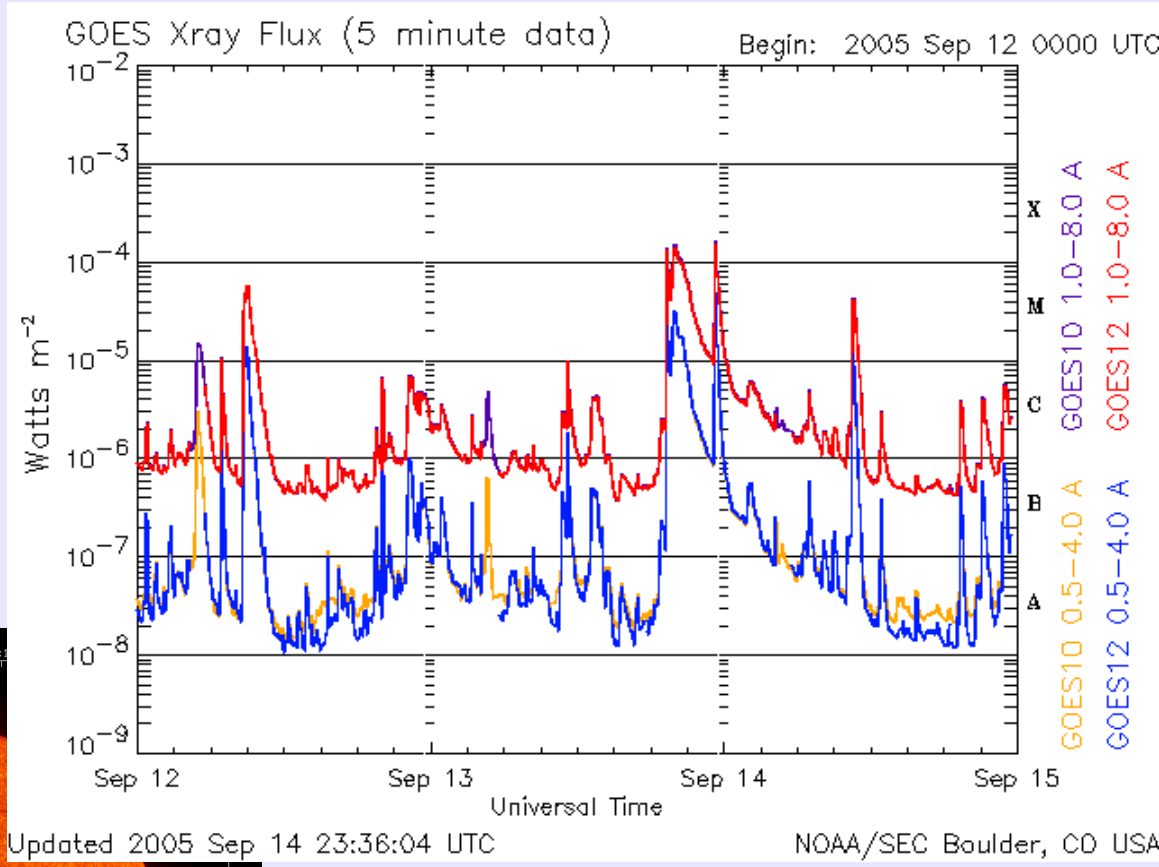
Solar Flares: what?

- Abrupt increase in radiative output *at all wavelengths*
  - 10^{28}—10^{32} erg over 10—1000 minutes.









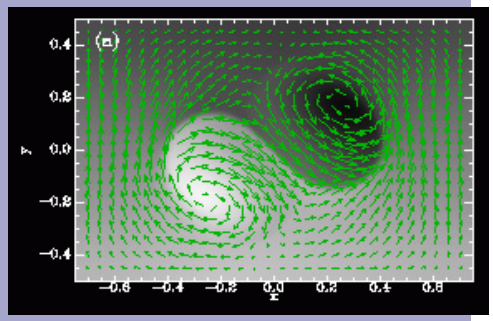
Solar Flares: why and who cares?

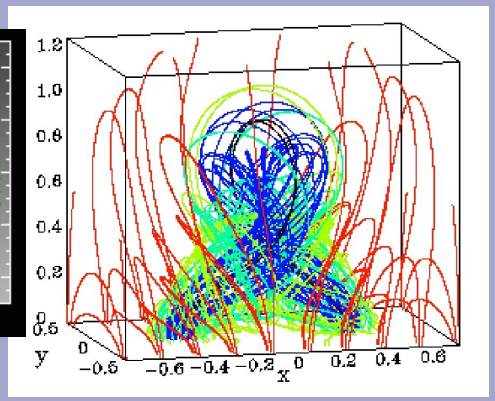
- **Unique role in solar physics research**
  - "The Sun would be a boring star if it had no magnetic field"
    - Solar Flares are the most dramatic examples of the magnetic field's influence.
  - Understanding solar flare production requires:
    - Understanding magnetic reconnection and other MHD processes such as the instability of magnetic modes (kink, tearing). Modeling these processes challenges present computational bounds.
    - Understanding the magnetic field and its interaction with the plasma in the "whole box", even though observations of the magnetic fields are only routine at the boundary.

- Unique role in "Space Weather"
  - Ionizing radiation can disrupt communications and pose radiation hazards
    - Time-of-flight governed by *c*
  - Associated Solar Energetic Particle events (if they occur) are deadly for astronauts.
  - True *forecasting* capabilities are required.

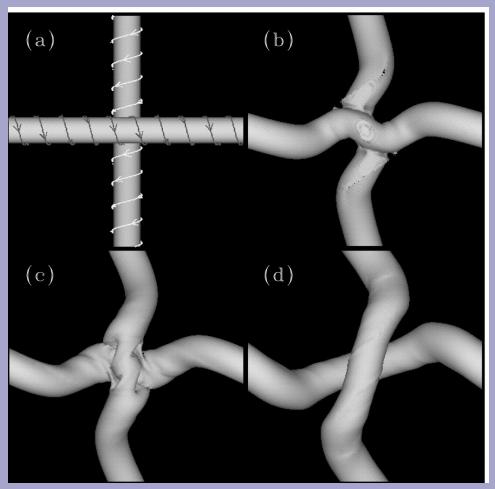
“How do I get a solar flare?”

Modeler's view:



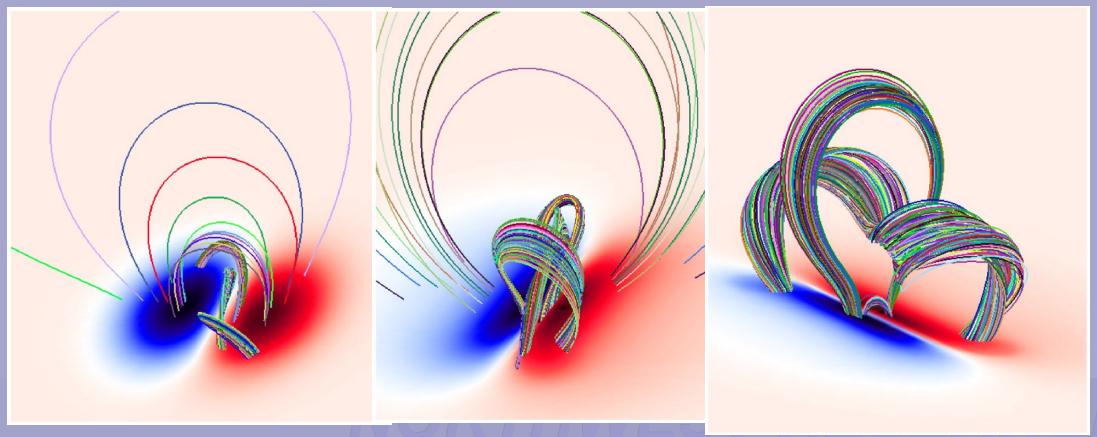


Fan & Gibson 2003, 2004

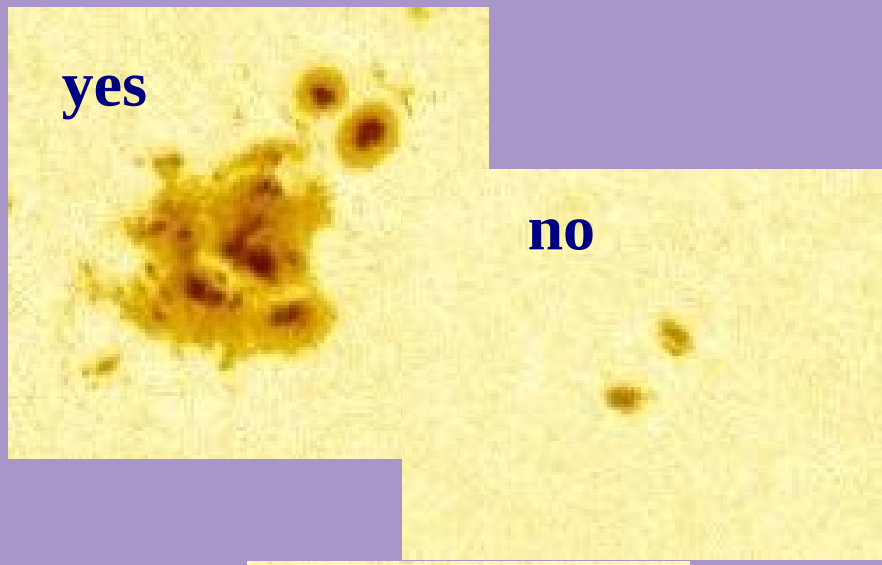


Linton & Antiochos 2002

Amari, Luciani, Aly, Mikic & Linker 2003



Forecaster's view:



Goals:

- *Test the null hypothesis:*

  "There is no measurable distinction between a flare-imminent and flare-quiet active region."
- If the null hypothesis is *rejected*, identify the *unique* signature of a "flare-imminent" active region.
- Develop a physics-based *objective* flare-forecasting approach.

General Considerations:

- Utilize appropriate data
  - Ensure sufficient data for statistical significance
  - Include data for "control" targets which did not flare
- Rely on objective calculations and statistics
  - Minimize "operator influence"
  - Account for flare "misses" as well as false alarms
- Wear two hats:
  - Balance and consider requirements of both flare forecaster and the numerical modeler.

Motivation for Using Magnetic Field Observations

- Energies released in solar flares consistent with magnetic energy in active regions, and inconsistent with other energy reserves.
- Magnetic reconnection in the corona is temporally consistent with needed trigger for the flare.
- Routine observations of the coronal magnetic field are not (yet)available, but photospheric magnetic field measurements are available.
- Various researchers have recently developed different ways to make flare forecasts from photospheric magnetic field measurements - how well do the forecasts do?

DATA: vector magnetic field maps of solar active regions

- **Imaging Vector Magnetograph, Mees Solar Observatory, Hawai`i**
- **Imaging Fabry-Perot system**
- **4 ' field-of-view, 1'' spatial resolution, 0.07nm spectral resolution**
  - polarization spectra sampled @ 30 positions across FeI 630.25nm line, g_L=2.5
  - Few-minute cadence
- **Full *B(x,y)* to characterize active region's magnetic field state.**

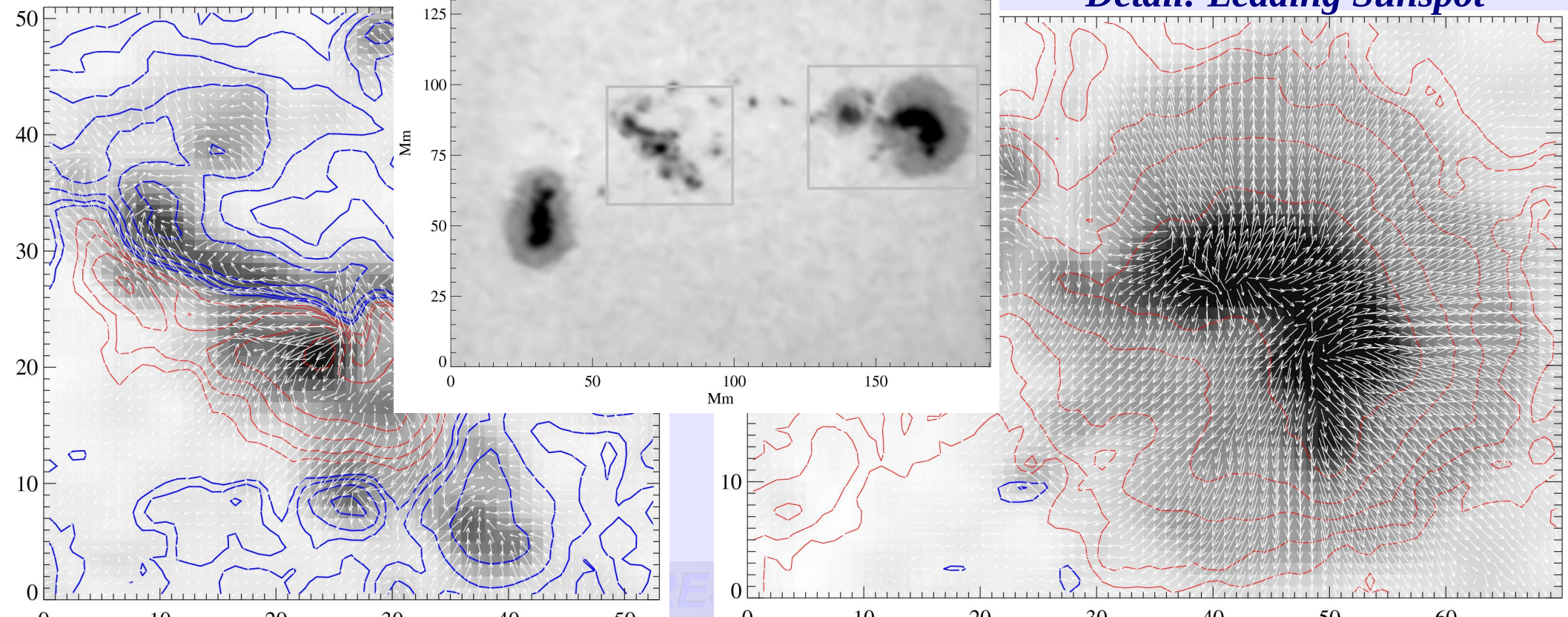
Detail: δ-type emerging sunspot

Full IVM field of view

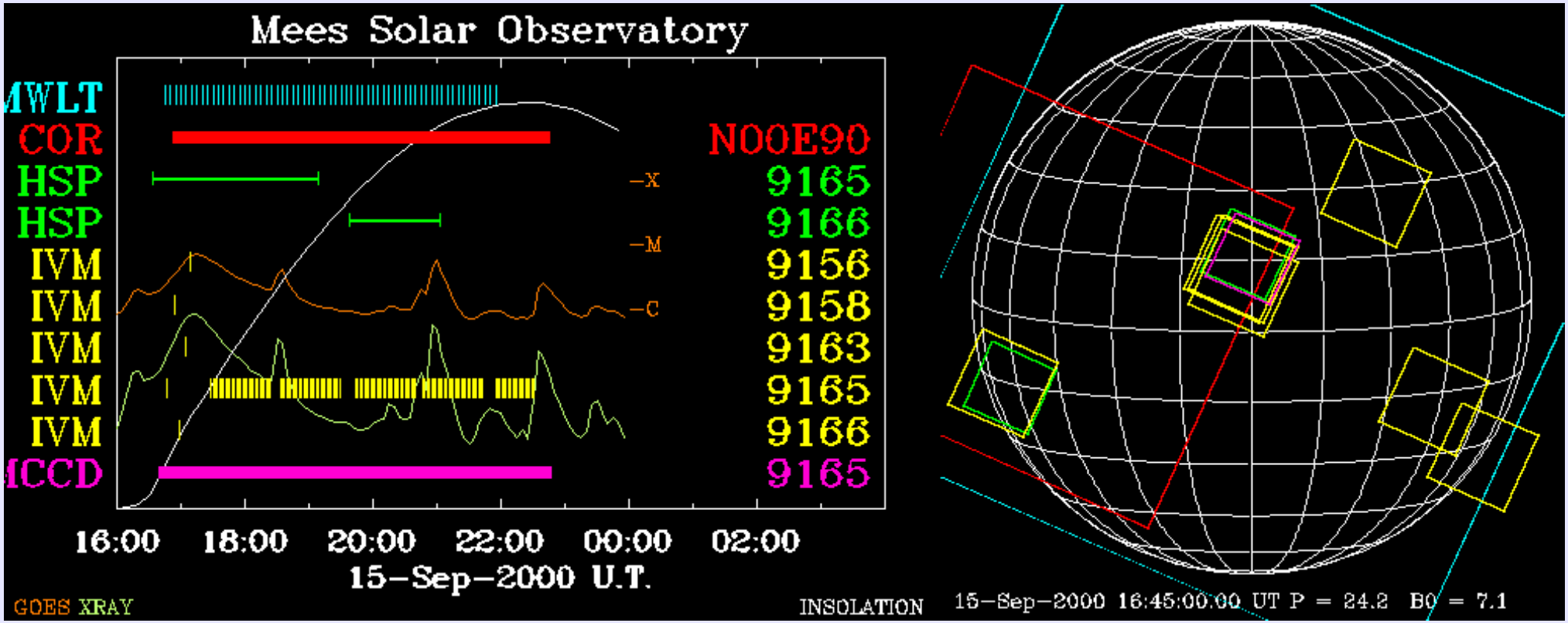
Red/Blue: contours of B_z

Arrows: B_{horiz}

Detail: Leading Sunspot



- IVM's observing sequence:
  - “Survey” magnetograms (single magnetograms of every visible active region)
  - Time-series observations of target region (chosen by Max Millenium program or similar campaign).
- Synoptic operation 1992-2006; large data base available.
  - 1,000+ daily survery magnetograms of NOAA regions 2001--2004

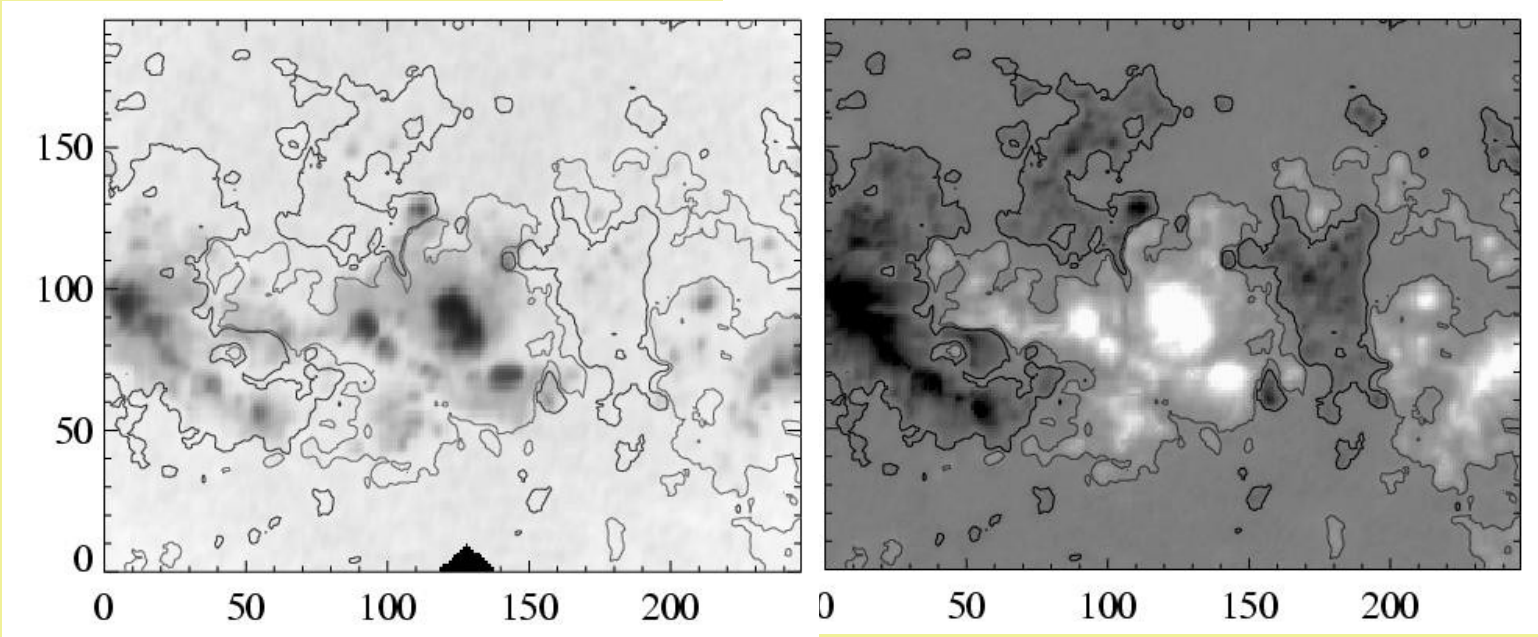


Data Analysis:

From *B(x,y)*, describe the distribution, morphology, and complexity of the photospheric magnetic field:

- **the magnetic field strength |B| and direction φ, γ**
- **the horizontal gradients of the magnetic fields** $|\nabla_h B| = \sqrt{(\partial B/\partial x)^2 + (\partial B/\partial y)^2}$
- **the vertical current density** $J_z = \nabla_h \times B_h \propto \partial B_y/\partial x - \partial B_x/\partial y$
- **the magnetic twist and current helicity density** $\alpha = J_z/B_z,\ h_c = B_z J_z$
- **the shear angle (deviation from potential)** $\Psi = \cos^{-1}(\vec{B}^p \cdot \vec{B}^o / B^p B^p)$
- **Proxy for the magnetic free energy** $\rho_e = (\vec{B}^p - \vec{B}^o)^2/8\pi$

IVM data of NOAA AR10030, 15 July 2001. *Left*: continuum image, scale is in Mm; *Right*: Radial component of the magnetic field, positive/negative (white/black).
For each "pixel", the full magnetic vector ***B*** is measured.



Parameterization: objective "single number" descriptors

- ***Moment analysis:***

$$mean\ \bar{x} = \frac{1}{n}\sum_i x_i$$

Average of distribution

$$standard\ deviation\ \sigma = \left[\frac{1}{n}\sum_i (x_i - \bar{x})^2\right]^{\frac{1}{2}}$$

Width of distribution

$$skew\ \varsigma = \frac{1}{n}\sum_i \left[\frac{x_i - \bar{x}}{\sigma}\right]^3$$

Asymmetry of distribution

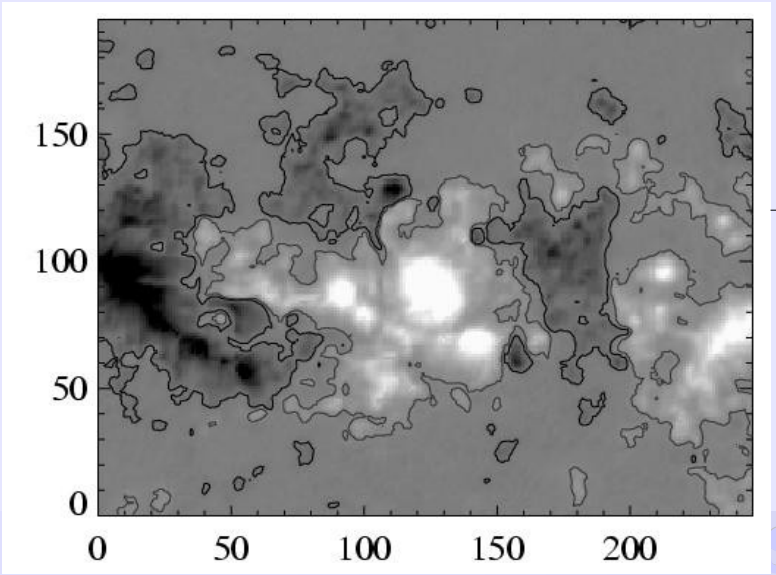
$$kurtosis\ \kappa = \frac{1}{n}\sum_i \left[\frac{x_i - \bar{x}}{\sigma}\right]^4 - 3.0$$

"flatness" compared to a Gaussian distribution

$$Total(x) = \sum_i x_i$$

Totals also included as appropriate

B(x,y)



→ $J_z(x,y,t)$, $h_c(x,y,t)$ → ~100 parameters describing the active region.

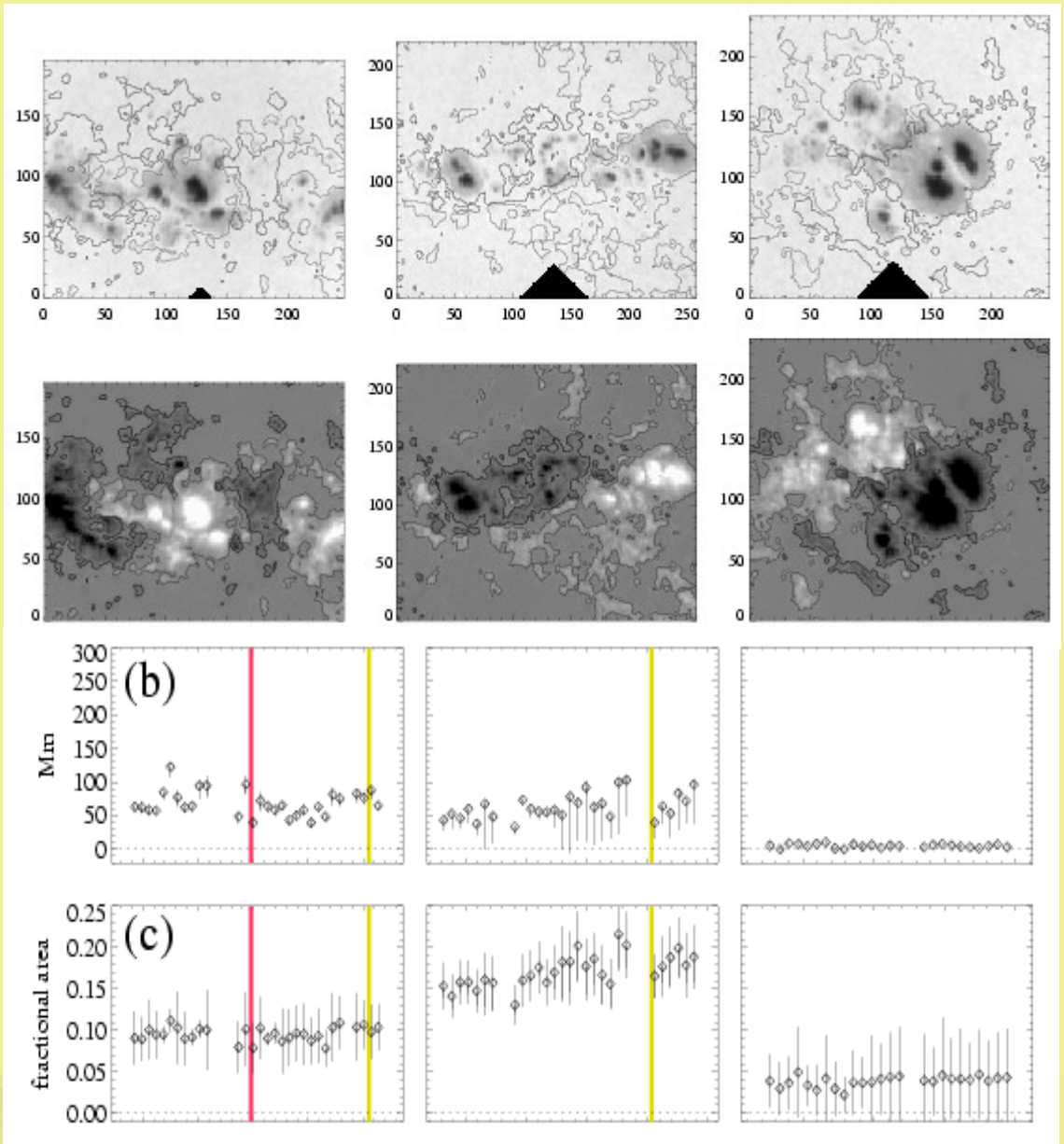
Leka & Barnes 2003 a,b

First Approach: ~1hr prediction windows

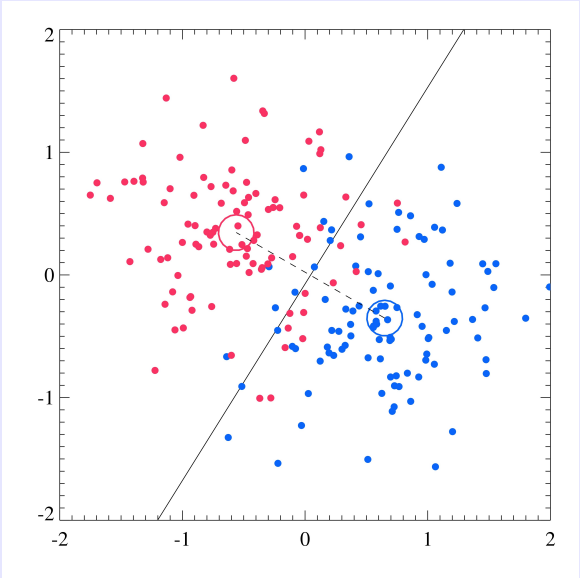
- Time-series of ***B***
- 7 Active Regions, all with moderate—high flare activity
- 10 “flare” *vs*. 14 “flare-quiet” epochs
- Full consideration of uncertainties, including “seeing”

Top rows: continuum and Bz of NOAA AR10030, 08636, and 08891.

Bottom rows: two example parameters derived from the magnetic shear angle ψ.
Top: length of magnetic neutral line with ψ ≥ 80°. Bottom: fractional area of the entire active region with ψ ≥ 80°. The start times for GOES X-class and M-class flares are indicated by vertical bars.



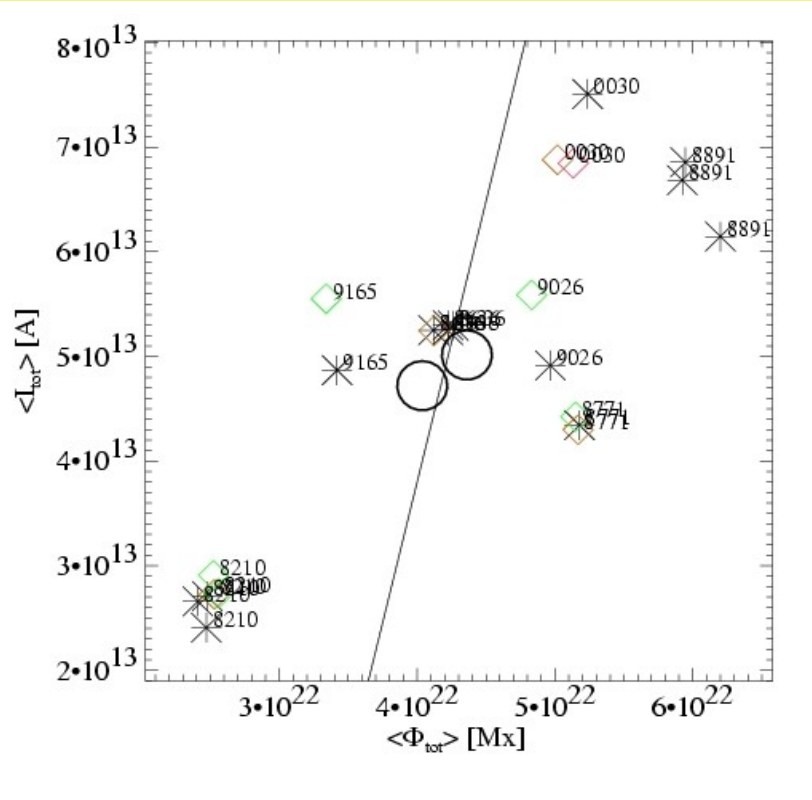
Discriminant Analysis



- Data sample two known populations (*e.g.*, flaring vs. flare-quiet)
- A "Discriminant function" best separates the samples:

$$f(x_1, x_2, \ldots x_n) = a_0 + a_1 x_1 + \ldots a_n x_n$$

- Statistical evaluation of whether samples originate from different populations via Hotelling's T^2 test
  - Power of variable x_n for "prediction" described by $|a_n|$
  - A new observation is "classified" as one/other population according to its observed $x_1, \ldots x_n$, and where it falls



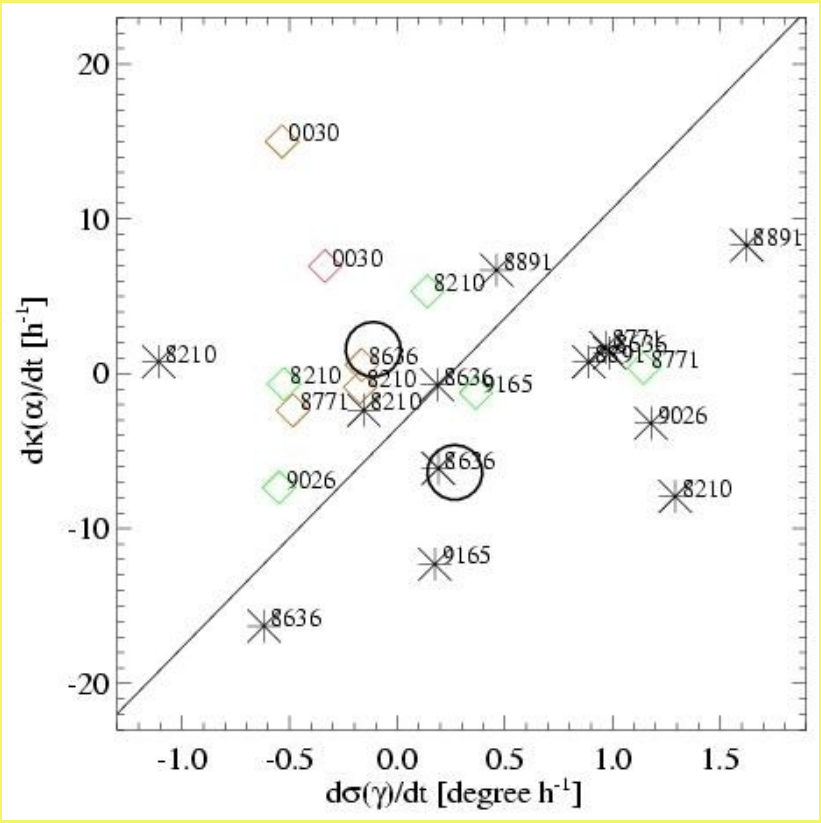
Example 1: Discriminant function from the total magnetic flux Φ_{tot} and the total electric current I_{tot} in active regions:

$$f = 0.0052 - 0.289 \langle \Phi_{tot} \rangle + 0.067 \langle I_{tot} \rangle$$

Graph: DF and the data for Flare-quiet (*) and flaring (◊,◊,◊ for C, M, X flares) points, plus the distribution means (O) are shown.

T^2 probability that the samples originate from different populations: 0.33.

Observed	*Predicted*		
		flare	no flare
	flare	5	5
	no flare	8	6



Example 2: Better performing and *completely* unintuitive:

$$f = 0.115 - 1.312\, d(\sigma(\gamma))/dt + 1.434\, d(\kappa(\alpha))/dt$$

T^2 probability that the samples originate from different populations: 0.94.

Observed \ *Predicted*	flare	no flare
flare	8	2
no flare	4	10

- ***Multiple Variables can be included simultaneously.***
  - 6-variable example:

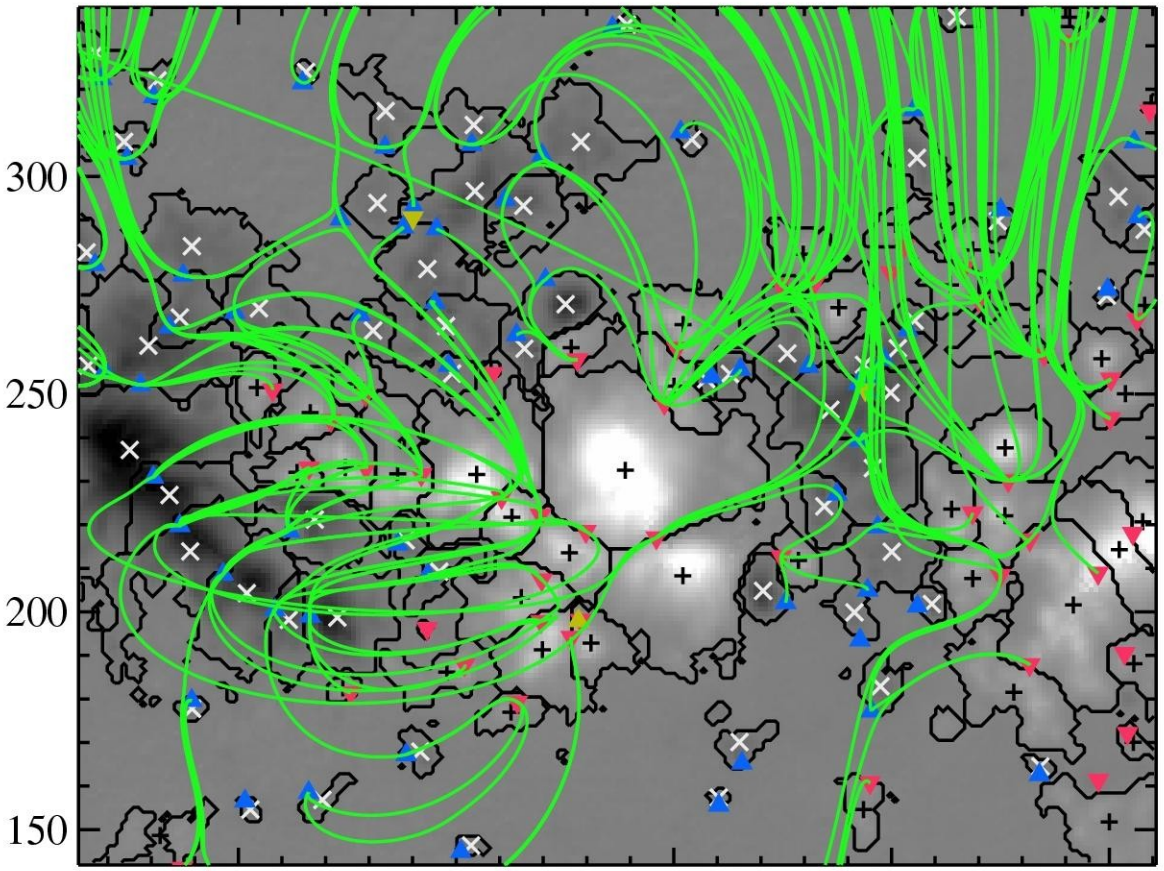
$$f = 1.021 - 11.098\,\langle\sigma(B_h)\rangle + 7.460\,d\overline{B_z}/dt + 8.330\,\langle\varsigma(J_z^h)\rangle - 3.829\,\langle\kappa(J_z^h)\rangle - 7.718\,\langle A(\psi > 80^\circ)\rangle - 3.834\,d|\alpha_{ff}|/dt$$

Observed \ *Predicted*	flare	no flare
flare	10	0
no flare	0	14

  - T^2 Probability: 0.9999
- Statistical flukes are likely: this is a *demonstration*.

The Coronal Complexity

- Magnetic reconnection believed to occur in the solar corona
  - Use photospheric ***B*** to investigate *coronal* ***B***
- **Magnetic Charge Topology** model
  - Partition the ***B*** maps, model as point sources, potential-field extrapolation, determine the coronal connectivity matrix
  - Characterize the coronal topology by the magnetic connectivity, distribution and character of magnetic nulls and separators.



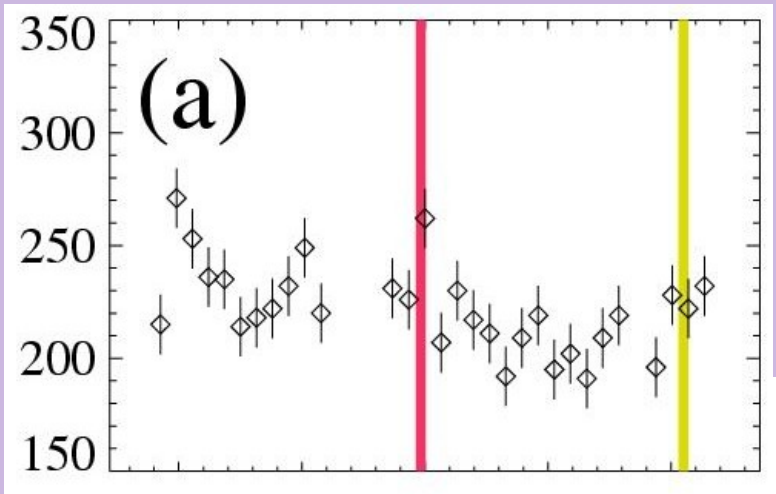
MCT analysis of NOAA AR10030 with 113 sources. "A" , "B" and upright null points (▼ , ▲ , ▽) and separator field lines (—). Axes are in megameters.

Barnes, Longcope & Leka 2005

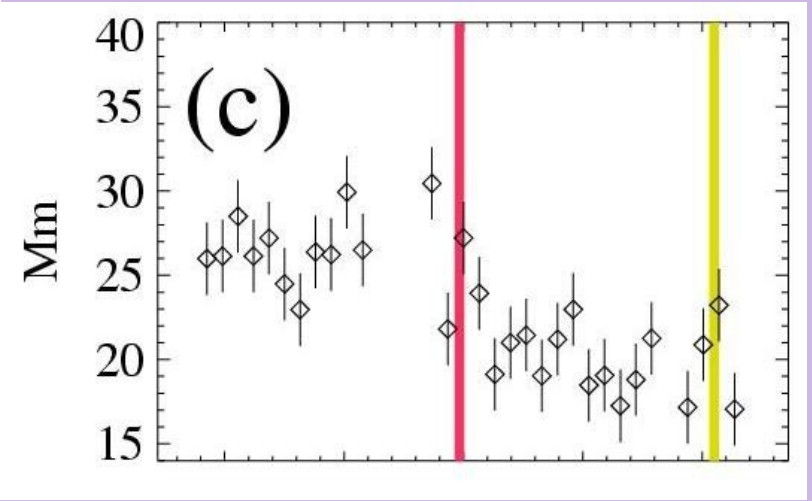
Parameterize: describe the topology map with single parameters.

Examples:

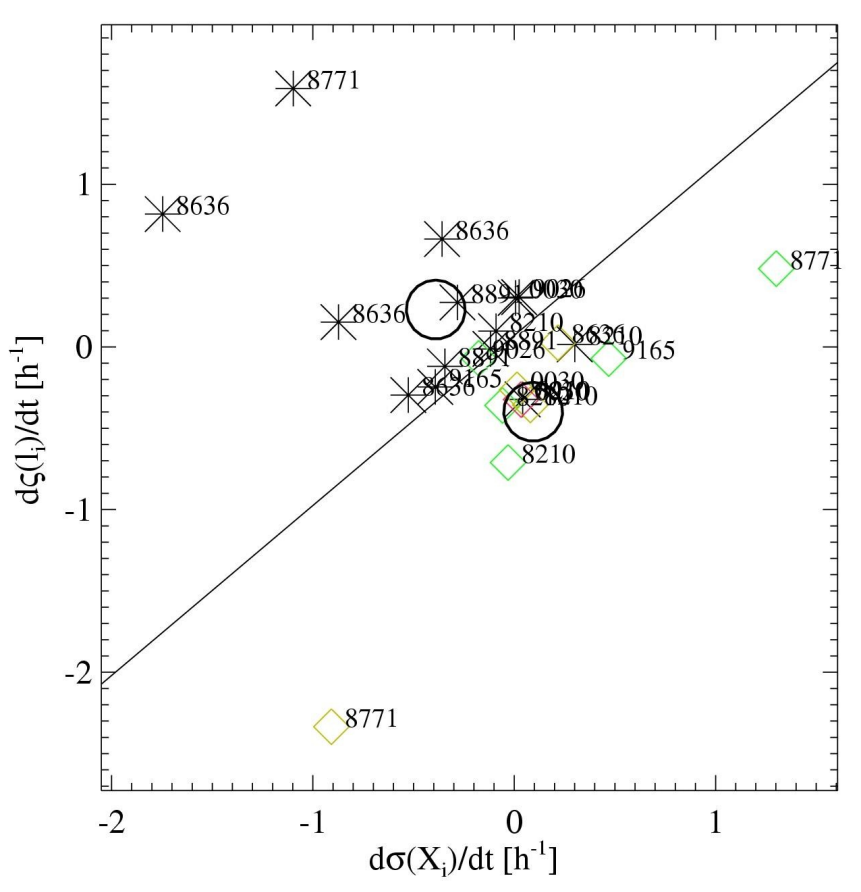
- mean, standard deviation of the flux in each connection: ψ_{ij}, $\sigma(\psi_{ij})$
- total flux weighted by the connection distances: $\varphi_{tot} = \Sigma\varphi_{ij} = \Sigma\psi_{ij}/|x_i - x_j|$
- Number of null points: $N = \Sigma N_p + \Sigma N_u$



For NOAA AR10030, (a) the total number of separators and (c) the mean height of the separators The start times for the X-class and M-class flares are indicated by vertical bars.



Discriminant Function Analysis of MCT model



Example: Two-variable DF for the temporal variation of the standard deviation of number of separators per null (x-axis) *vs.* the temporal variation of the skew of separator lengths (y-axis).

T^2 probability: 0.974

Observed		*Predicted* flare	no flare
	flare	9	1
	no flare	2	12

- ***Only 4 variables*** *now needed simultaneously for a "perfect" classification table:*

$$f = 0.26 - 1.02\frac{d}{dt}\varsigma(\psi_{ij}) + 1.80\frac{d}{dt}\sigma(X_i) - 1.62\frac{d}{dt}\varsigma(l_i) + 1.09\langle\sigma(r_{ij},\psi)\rangle$$

- MCT provides better probabilities overall of discriminating the populations:

the *Corona* may better indicate whether/when an active region will flare*.

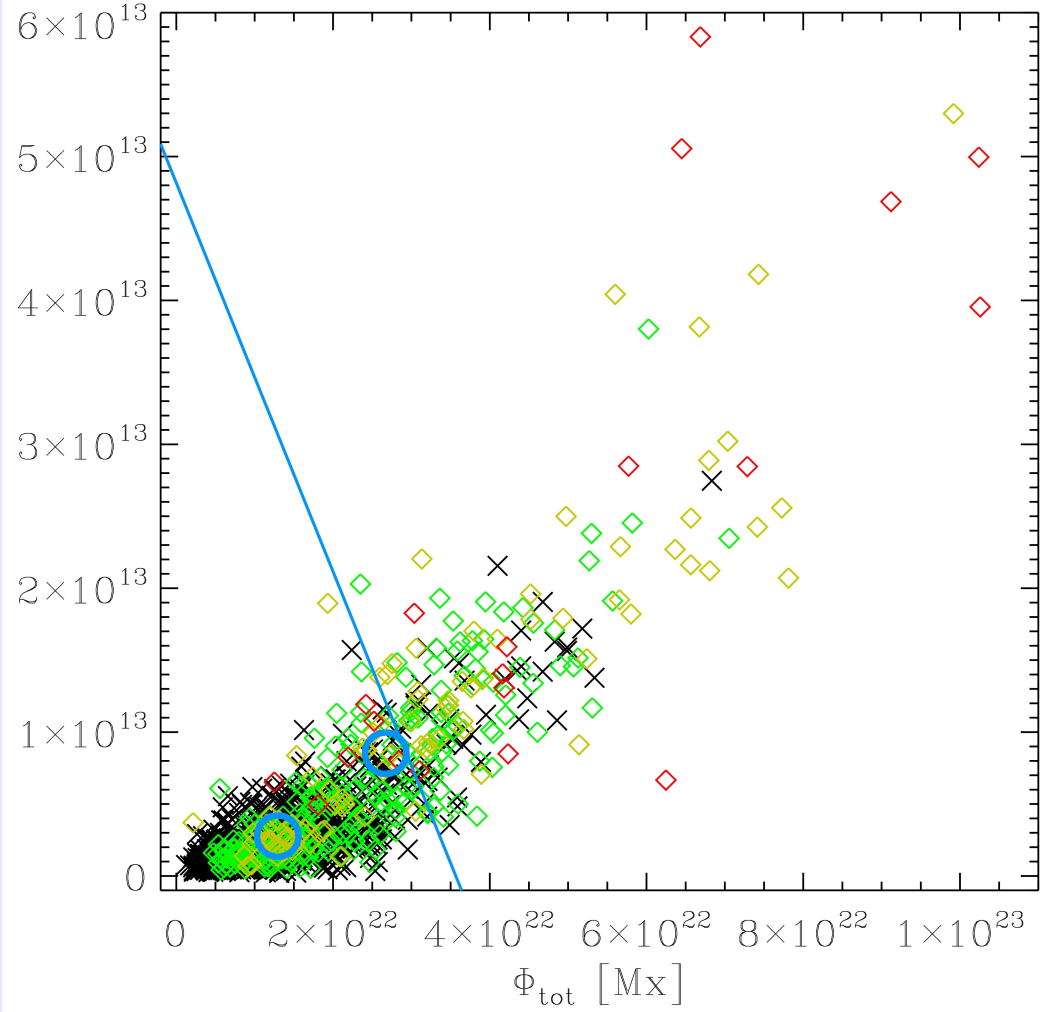
* statistically: *very* small sample!

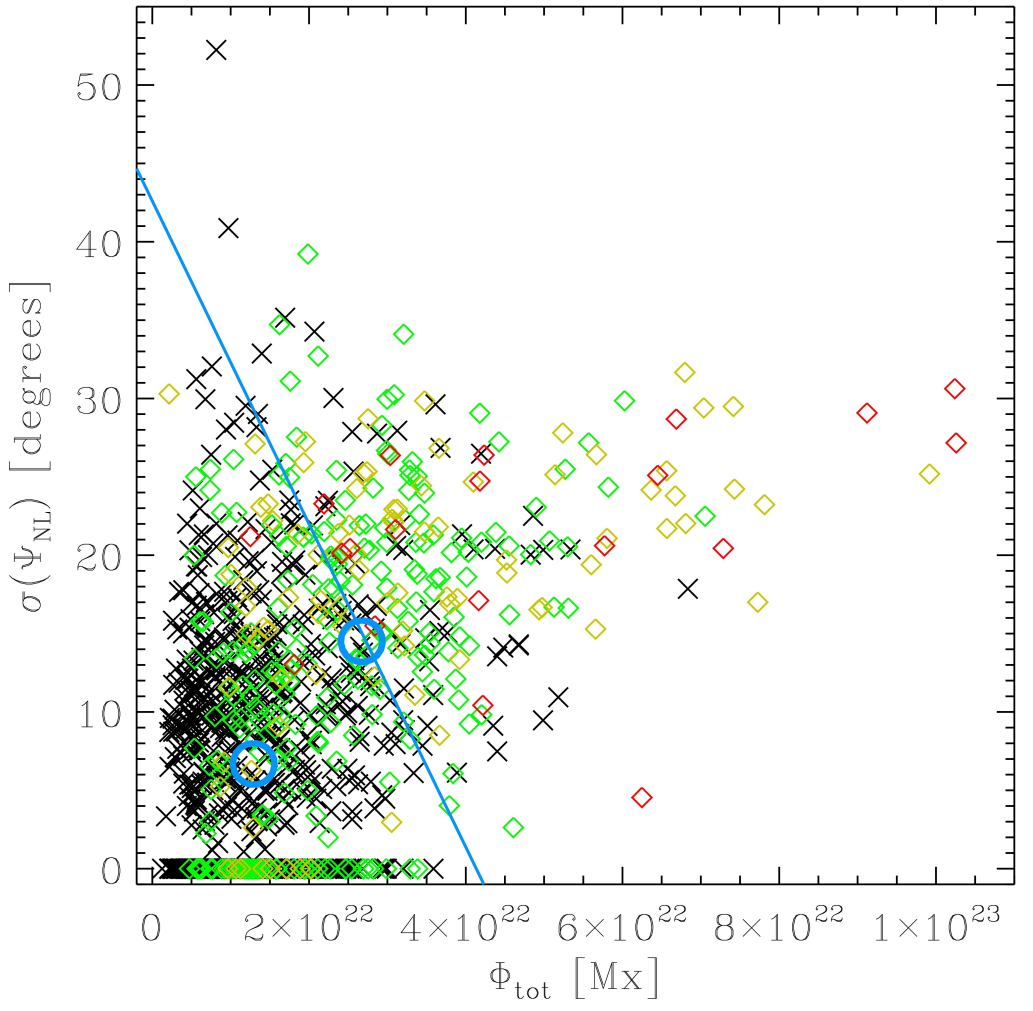
Quantifying, Objectifying *Daily* Flare Forecasts

- **Focus on *1-day* prediction window with IVM "survey data" archive**
  - Same parameterization of ***B*** morphology
    - Shear angle, flux, energy, etc.
    - No temporal evolution
  - Better statistics: > 1,100 data points (2001--2004)
  - More sophisticated implementation of Discriminant Analysis:
    - Unequal covariance matricies
    - Unequal population sizes

Event Definition:

- **GOES events within in 24-hr after magnetogram.**
  - **29.6% produced at least event ≥ C1.0,**
  - **9.2% produced at least one event ≥ M1.0, and**
  - **1.7% produced at least event ≥ X1.0**





Examples: 2-variable discriminant functions for ***(left)*** the strongly correlated pair Φ_{tot} , I_{tot} and ***(right)*** the uncorrelated pair Φ_{tot} ,$\sigma(\Psi_{NL})$. Non-flaring regions (x), and flaring regions (◊) are shown with the largest flare in any 24-hr period **(C, M, X)**, with the mean of each sample (**O**) and the discriminant function (—). There are a number of points with $\sigma(\Psi_{NL})$=0.0 from regions where there are no well measured horizontal fields on the neutral line. The points with Φ_{tot} ≥10^{23} Mx are NOAA AR 10486 (source of most Halloween Storms).

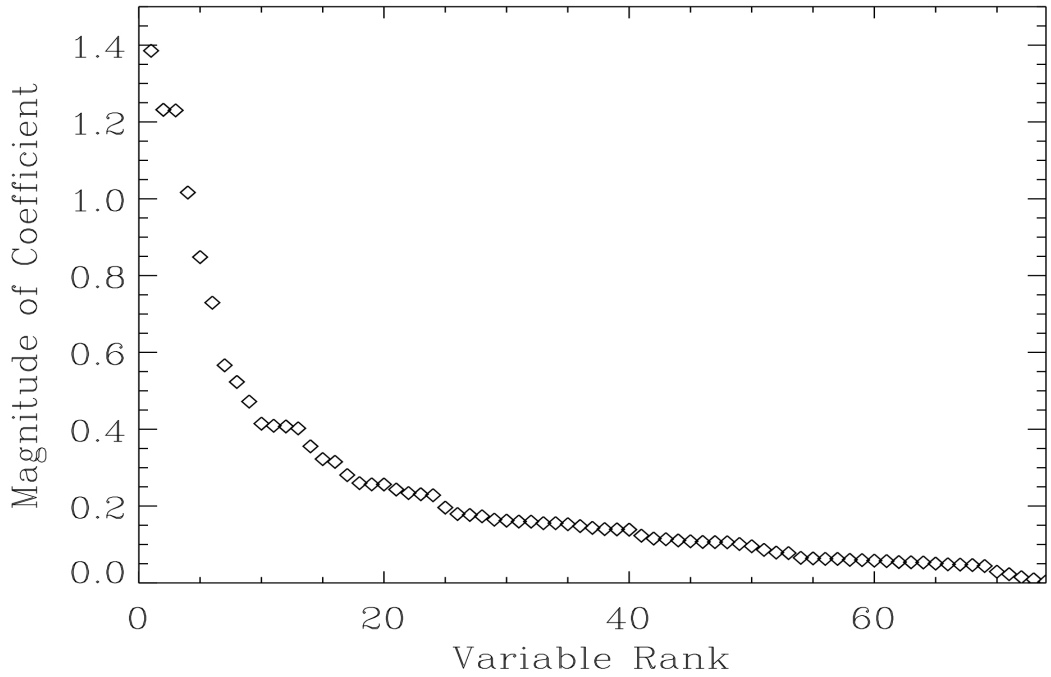
Which are the "Best Parameters"?

- **Short answer: none by itself.**
- **Discriminant analysis provides a *discriminant function*.**
  - **coefficients give the *relative predictive power:***

$$f_{all} = 1.385\Phi_{tot} + 1.232\, I_{tot} - 1.230\, I^{h}_{tot} - 1.016\, E_e + 0.848\, |\overline{\nabla Bz}| + \ldots$$

- **Many variables are correlated, reducing their "predictive power". Most predictive power obtained with first/best 10 variables *used together*.**
  - **There is no *single* "smoking gun".**
  - **Best approach: combine a selection of *uncorrelated* variables.**

Magnitudes of the coefficients in the all-variable discriminant function. There is a *rapid decrease* for the first several variables, after which the magnitudes decrease slowly, indicating that more variables provide little additional predictive capability.



Best variable combinations

Photospheric parameters, ≥ C1.0 flares

n	*variables*	*success rate*
0	predict nothing flares	0.704
1	Φ_{tot}	0.772
2	Φ_{tot}, $\sigma(\Psi_{NL})$	0.794
3	Φ_{tot}, $\sigma(\Psi_{NL})$, $<\Psi>$	0.803
4	Φ_{tot}, $\sigma(\Psi_{NL})$, $\sigma(\alpha)$, $\sigma(\gamma)$	0.804
74	*all*	*0.810*

Classification Table

All photospheric variables,
≥ C1.0 flares:

		observed	
		Event	No Event
predicted	Event	199	160
	No Event	70	783

So,

- There ***is*** a *small* difference detectable between flaring active regions and flare-quiet regions.
- ***How do we compare this to other methods?***
- ***How do we improve it?***

Probabilistic Forecasting

Adapt the discriminant analysis to provide *probabilities* that a measurement belongs to either group. Assume the *a priori* probability of membership in a population is proportional to the sample size. Then the probability that a measurement x belongs to a flaring region is given by

$$P_f(x) = \frac{n_f f_f(x)}{n_f f_f(x) + n_q f_q(x)}$$

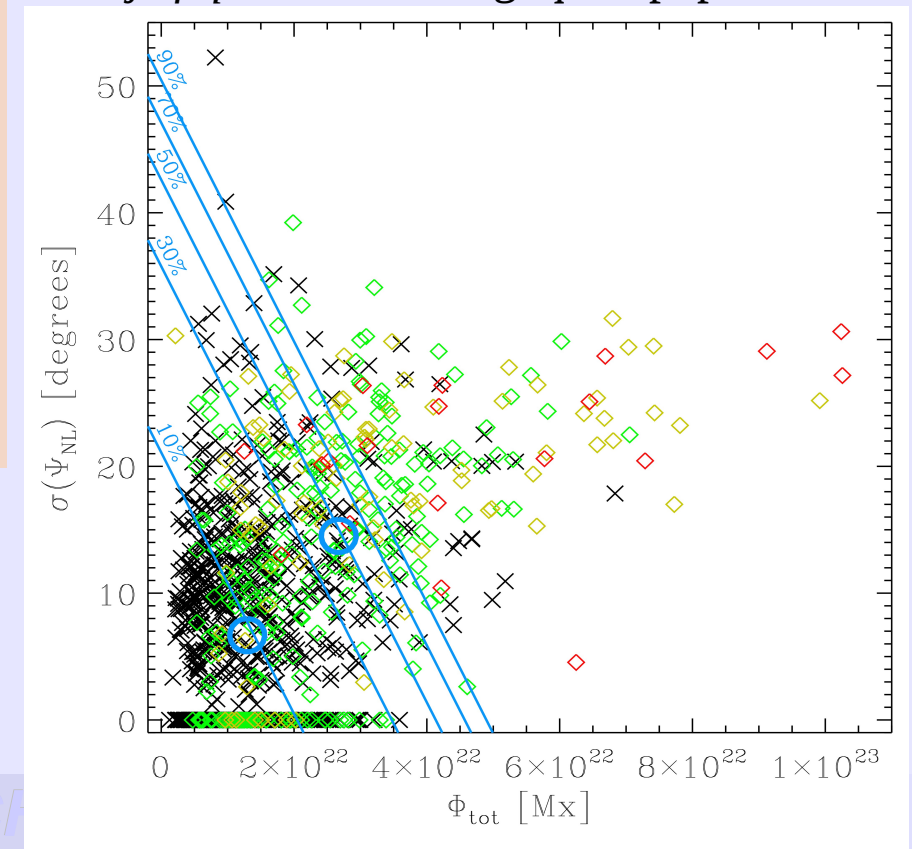
Where n_j is the sample size, $f_j(x)$ is the probability density function for population j, and $j=f,q$ refer to flaring, quiet populations.

Multivariate Gaussian probability distribution of p-dimensions is given by:

$$f_j(x) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{p/2}} \exp\left[\frac{-1}{2}(x-\mu^{(j)})' \Sigma^{-1}(x-\mu^{(j)})\right]$$

Where $\mu(j)$ is the vector of mean parameter values, Σ is the population covariance matrix, x is the vector of parameter values for the new active region to be classified.

Substituting this and expressions for the covariance matrices into the above $\boldsymbol{P_f(x)}$ provides the ***forecast probability*** at any point.



Verification Statistics: *for direct comparison to other methods*

<f> : *average over all data points of the forecast probability*

<x>: *average " " " " of the observation*
(x = 0 if quiet, x=1 if flaring).
If <x> is lower than <f>, a method <u>over</u>-predicts activity.

<f|x=1> : *average of the forecast probability for flaring data,* ***best is 1.0***
<f|x=0> : *" " " " quiet data,* ***best is 0.0***

MAE (f, x) : *mean absolute error,* ***best is 0.0***

SS(f, x) : *climatological skill score,*
value added over "default" or "no data" probability forecast.
best is 1.0,
SS=0.0: no value added,
SS < 0.0: worse than "no data"

Success Rates, Flare Threshold and

"if it sounds too good to be true..."

Photospheric all-variable success rates:

- **81% for ≥ C1.0**
- **93% for ≥ M1.0**
- **98.5% for ≥ X1.0**

Looks promising, especially for X-class flares!

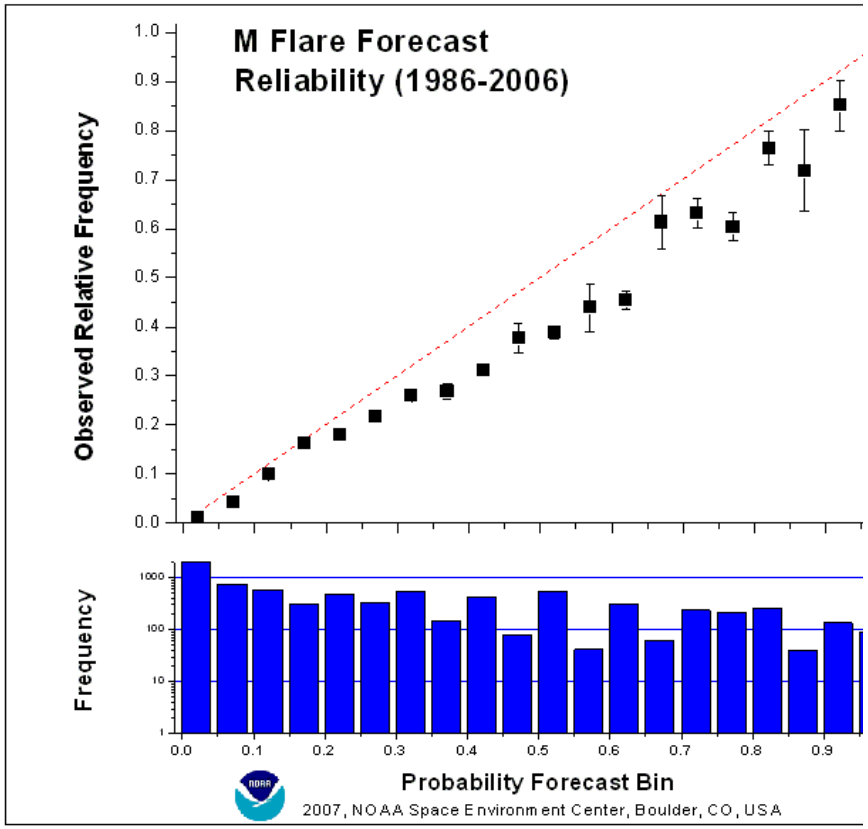
	Discriminant Analysis			Bayesian (M. Wheatland)		NOAA/SEC	
	(≥C1.0)	(≥M1.0)	(≥X1.0)	(M only)	(X only)	(M only)	(X only)
<f>	0.279	0.076	0.013	0.294	0.040	0.298	0.064
<x>	0.296	0.092	0.017	0.262	0.035	0.262	0.035
<f|x=1>	0.544	0.422	0.450	0.510	0.122	0.551	0.244
<f|x=0 >	0.167	0.041	0.006	0.217	0.037	0.208	0.057
MAE(f,x)	0.253	0.090	0.015	0.289	0.066	0.271	0.081
SS(f,x)	0.346	0.252	0.123	0.258	0.078	0.262	-0.006

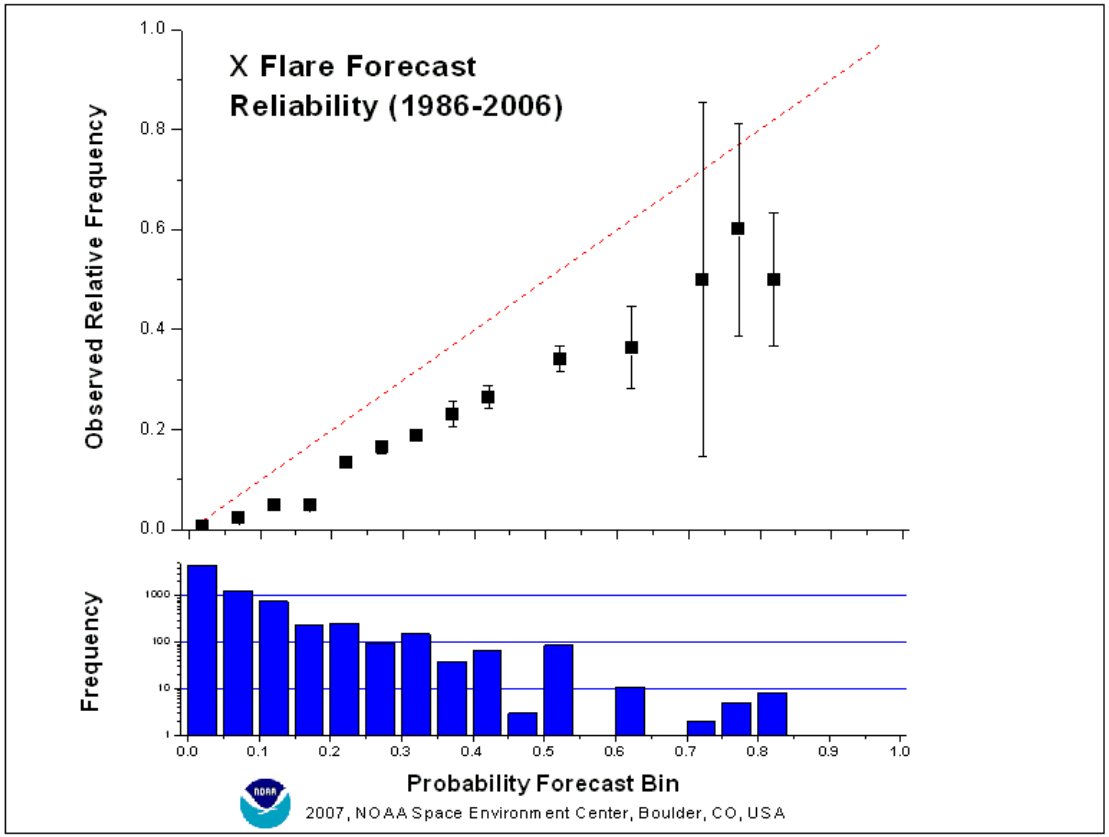
***The truth*: little value added from a "no-data" forecast**

***Also true*: DA +*B* is *comparable to established NOAA/SEC forecasts*.**

Reliability Plots

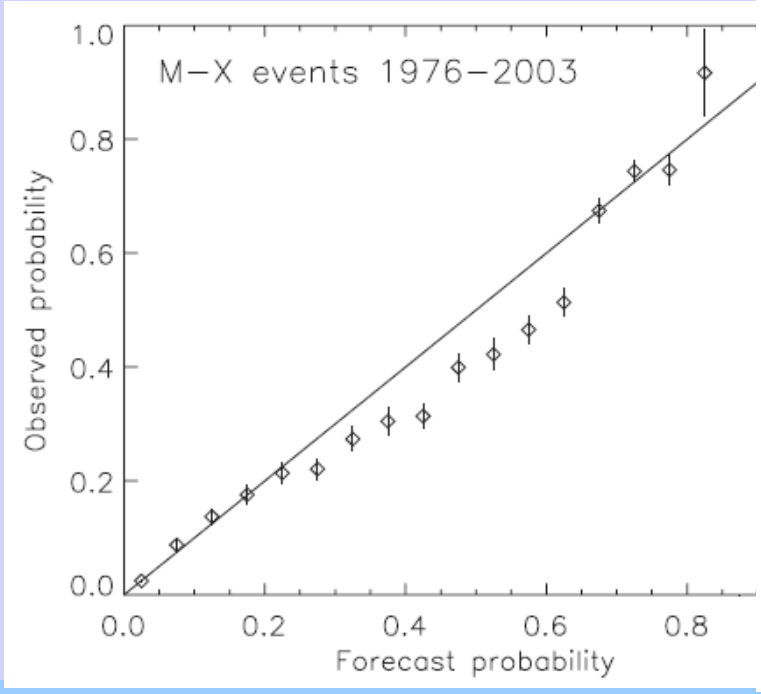
- **Perfect forecast: all points lie on the line.**
- **Note small error bars with larger sample sizes and *vice versa*.**
- **"under" the line: method over-predicts probabilities and *vice versa*.**

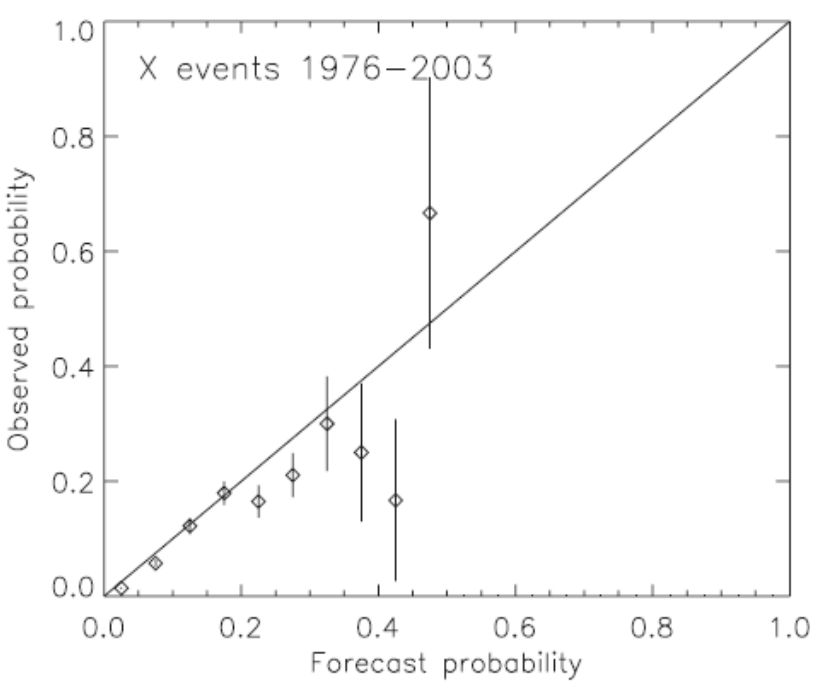




http://www.swpc.noaa.gov/forecast_verification/index.html

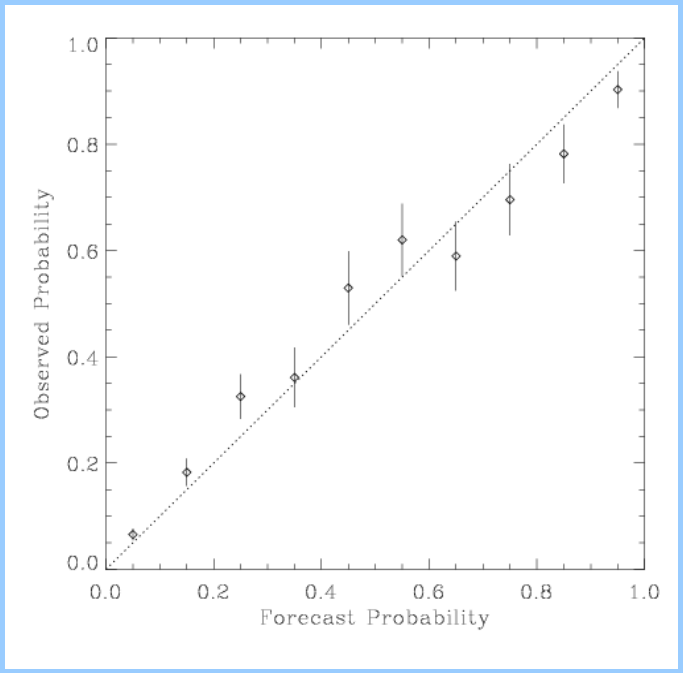
Reliability plots for flare persistance & Bayesian statistics (Wheatland 2005)



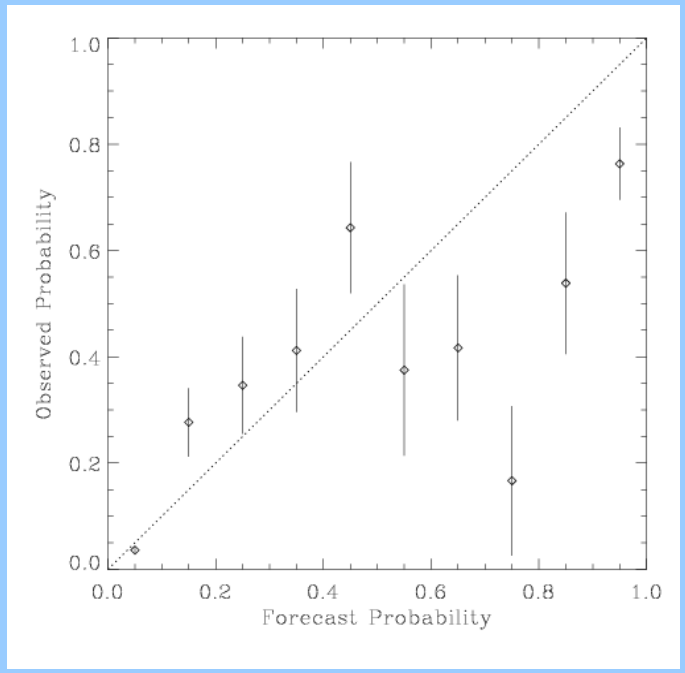


Reliability plots for Photospheric Parameters & Discriminant Analysis:

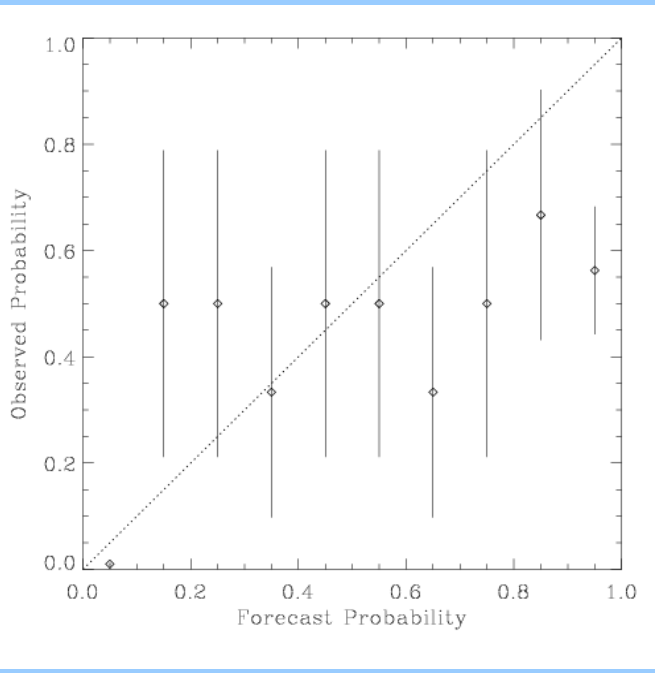
≥C1.0



≥M1.0



≥X1.0

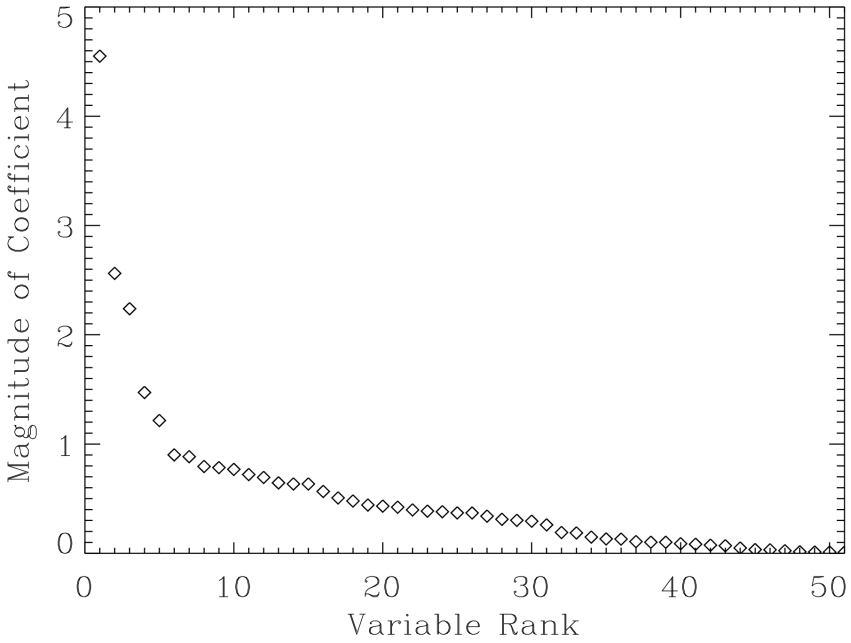


Coronal MCT parameter Discriminant Analysis

Classification Table

All variables, ≥ C1.0 flares:

predicted	*observed* Event	No Event
Event	190	169
No Event	66	787



- **Same "flare, quiet" groups**
- **different variables**
  - **derived from MCT/corona, rather than photospheric parameters.**
- **Same basic DA approach**

Best variable combinations coronal parameters, ≥ C1.0 flares

n	***variables***	***success rate***
0	predict nothing flares	0.704
1	φ_{tot}	0.770
2	φ_{tot}, $\sigma(\xi_{ij},\psi)$	0.784
3	φ_{tot}, $\sigma(\xi_{ij},\psi)$, $\kappa(\varphi_{ij})$	0.788
4	φ_{tot}, $\varsigma(\varphi_{ij})$, $\kappa(\varphi_{ij})$, $\overline{\psi}_{ij}$	0.793
47	all	0.806

Same behavior as photospheric analysis for multiple simultaneous variables

Coronal MCT all-variable success rates:

- **81% for ≥ C1.0 (default forecast: 70%)**
- **92% for ≥ M1.0 (default forecast: 91%)**
- **98% for ≥ X1.0 (default forecast: 98%)**

Here we go again...

	Photospheric Parameter Discriminant Analysis			Coronal MCT Parameter Discriminant Analysis		
	(≥C1.0)	(≥M1.0)	(≥X1.0)	(≥C1.0)	(≥M1.0)	(≥X1.0)
<f>	0.279	0.076	0.013	0.278	0.081	0.024
<x>	0.296	0.092	0.017	0.296	0.092	0.016
<f|x=1>	0.544	0.422	0.450	0.524	0.421	0.498
<f|x=0 >	0.167	0.041	0.006	0.174	0.046	0.016
MAE(f,x)	0.253	0.090	0.015	0.263	0.095	0.024
SS(f,x)	0.346	0.252	0.123	0.340	0.228	-0.262

- ***Still, little (if any!) value added*** **from a "no-data" default forecast.**

Non-Parametric Discriminant Analysis

Goal: Remove assumptions of Gaussian distributions with equal covariance matrices.

Approach: Estimate the probability density using a non-parametric technique

The Kernel Method for Nonparametric Density Estimation

Probability density is estimated by summing over the contribution from each data point, weighted by a given kernel function:

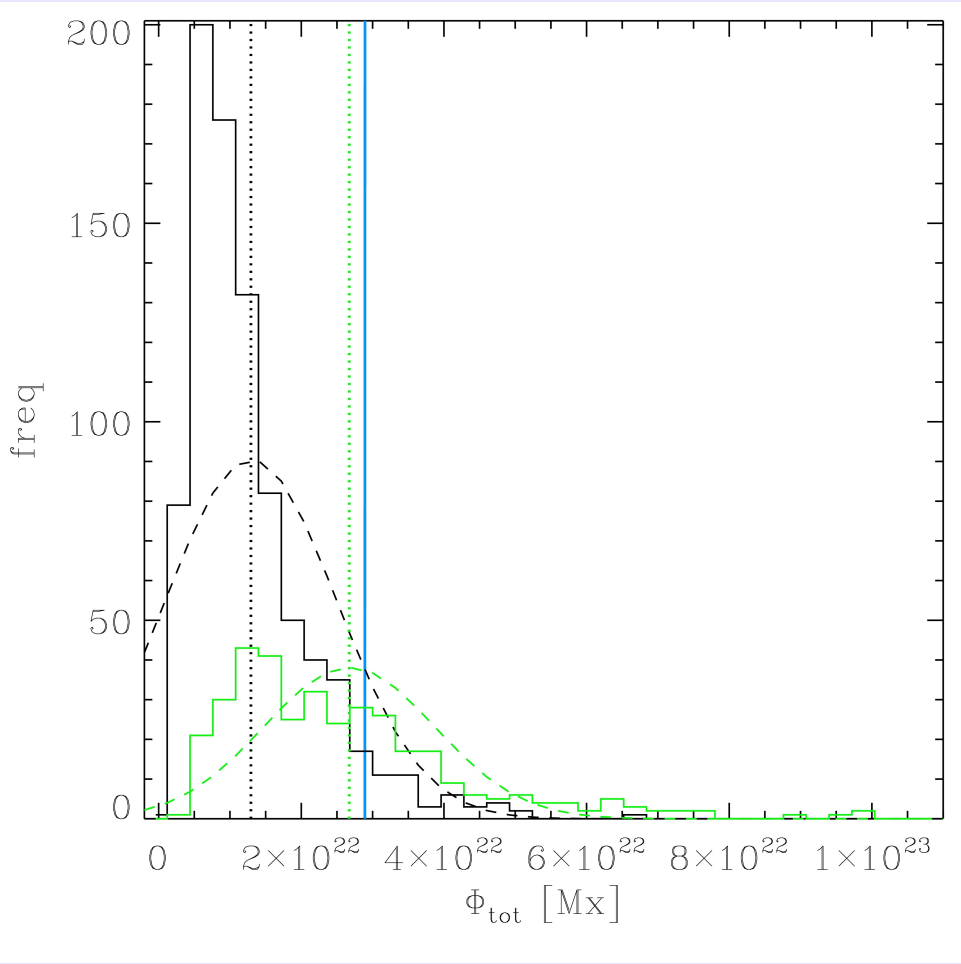
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(t)$$

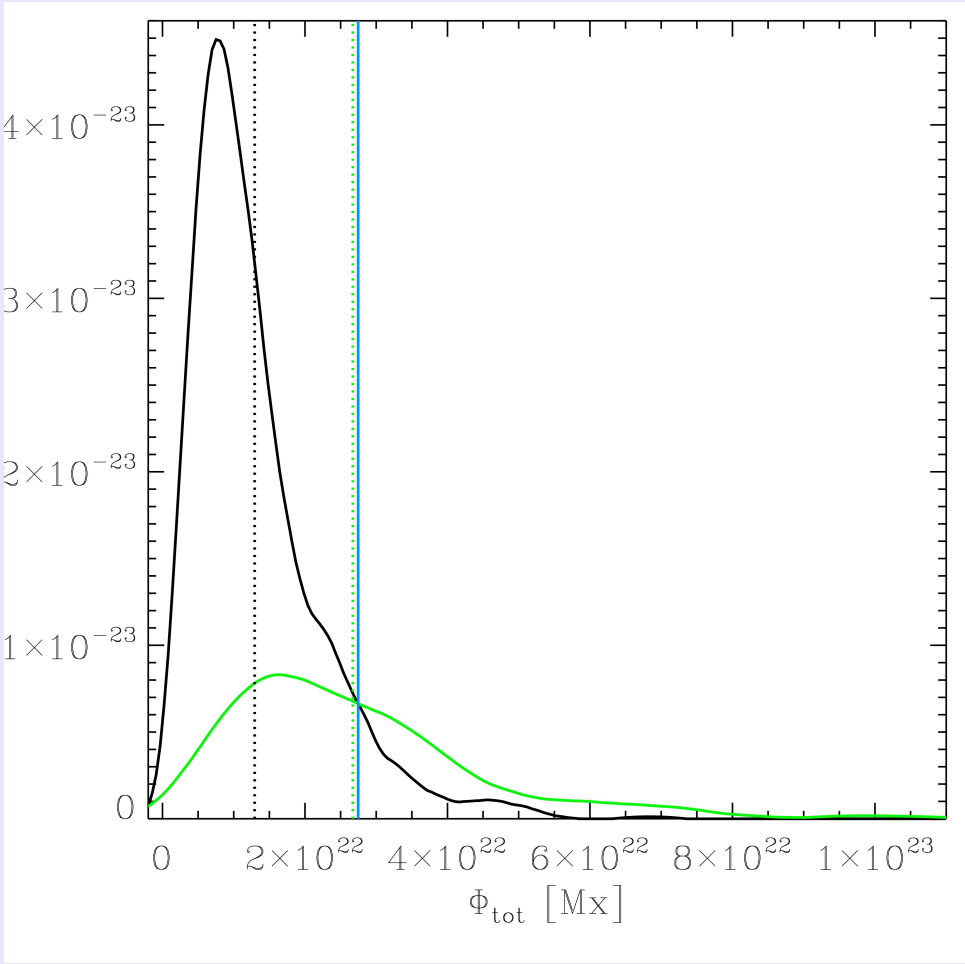
where $t=(x - x\_i)/h$, $K(t)$ is the kernel function, and h is a "smoothing parameter".

The most efficient kernel for univariate data, used here, is the Epanechnikov kernel:

$$K_e(t) = \begin{cases} \frac{3}{4\,\mathrm{sqrt}\,5}\left(1 - \frac{t^2}{5}\right) & |t| < 5 \\ 0 & \textit{otherwise} \end{cases}$$

A large h can oversmooth the probability density and hide real structure, but a small h may undersmooth the probability density, and keep small-scale but insignificant structure. Here, $h = \sigma\, n^{-1/5}$, where σ is the sample standard deviation.





Linear *vs* nonparametric discriminant functions for total flux Φ_{tot}. The **flaring probability density** and **non-flaring probability density** are **equal** at the discriminant boundary. Note the difference in ability to separate the distribution tails.

The nonparametric discriminant results in a higher correct classification rate, but in this case the improvement is small: 0.7739 compared to the linear discriminant rate of 0.7723.

Concluding Remarks:

Demonstrate the required approach:

- Quantitative analysis including flare-quiet examples as control data
- Appropriate statistical analysis.

Results helpful for many efforts:

- *Produce* physics-based objective solar flare prediction using observations of the solar magnetic field.
- *Guide* modeling efforts, as to the necessary boundary conditions and coronal topology for flaring active regions.

Results are mixed:

- *No single parameter* produces a unique pre-flare signal.
- Parameter *combinations* can do as well as NOAA/SWPC forecasts & Baysian methods
- Magnetic field + Discriminant Analysis is completely objective
- *Still, not much better than predicting that nothing flares.*

Significant room for improvement:

- Combine parameters with flare persistence, evolution, topology

Upcoming: *"Toward the Operational All-Clear Forecast"*, ***work***shop hosted by NOAA/SWPC, NWRA, and NASA/Space Radiation Analysis Group, April 2009.

Research and Method demonstrate value of long-duration observing facilities, preparing for HMI on the Solar Dynamics Observatory