

Limits on the Complexity of Empirical Models of Magnetic Storm Phenomena

Paul O'Brien

Fall AGU 2005, Paper SM11A-04

Accepted for publication in *Space Weather*

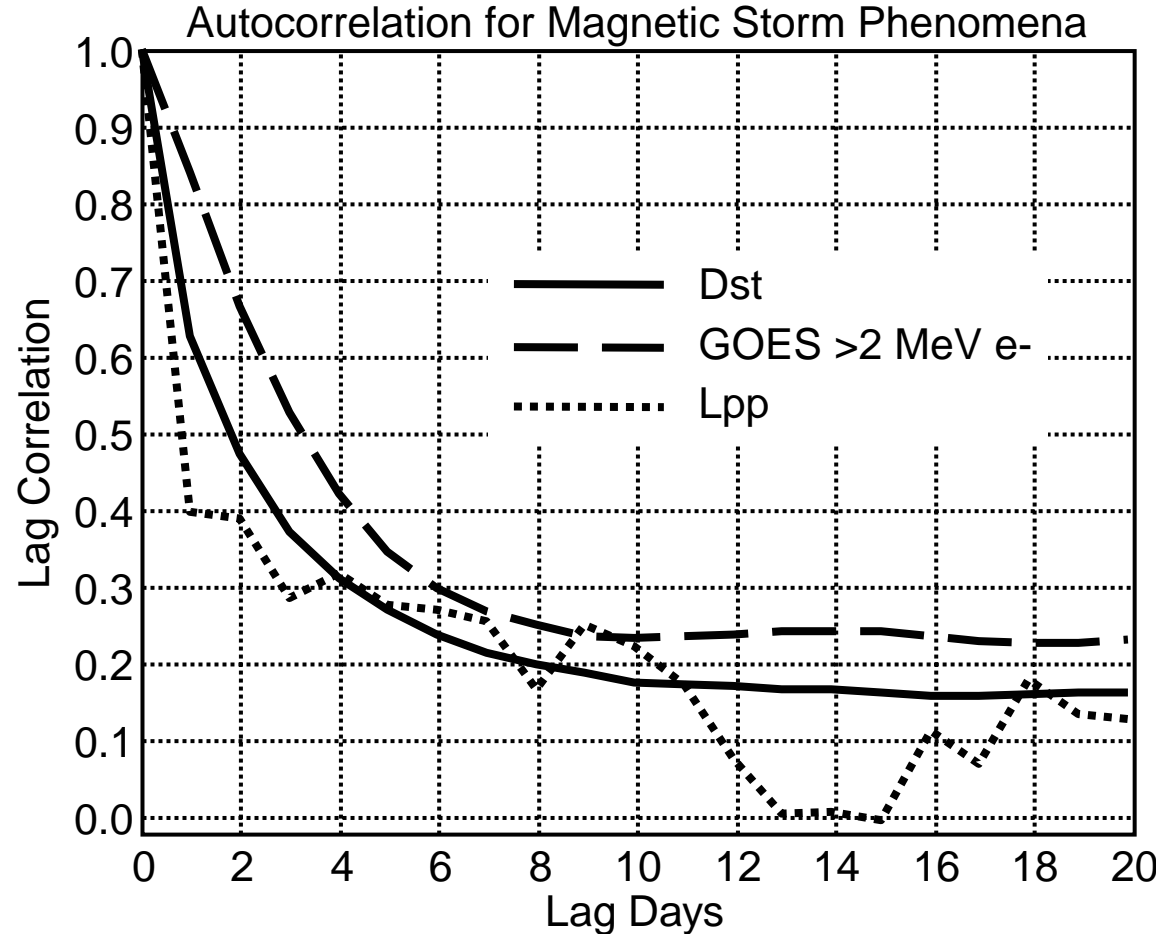
Research Funded in part by The Aerospace Corporation's Internal Research and Development Program

Introduction

- There are a growing number of empirical models of magnetic storm phenomena, particularly:
 - Dst
 - Relativistic Electrons at GEO
 - Plasmapause location (Lpp)
- These models are getting ever more complex, including more extra terms, and having more free parameters
- The more free parameters, the better the fit, but how good of a fit is required to justify n parameters?
- Auto correlation (aka serial correlation) in the observations severely decimates the number of independent samples of magnetospheric behavior from which to derive empirical models with many free parameters

Serial Correlation

- There are about 33 magnetic storms a year, depending on the definition
- Between these storms the inner magnetosphere approaches a quiet state
- The long intervals of quiescence provide little information on magnetospheric dynamics
- During the storms, everything responds at the same time in a highly correlated way
- The time series of Dst, >2 MeV e- at GOES, and CRRES Lpp all show significant serial correlation for over 10 days.
- We can only count each storm as 1 independent observation
- We only get ~33 samples/year, or ~361 samples/Solar cycle



An Example with Dst

- Let us pretend that we are building a simple linear model of Dst with several possible terms:

$$\text{Dst}(t+dt) = a \text{ Dst}(t) + b \text{ VBs}(t) + c \sqrt{P}(t)$$

- We begin by considering an even simpler model:

$$\text{Dst}(t+dt) = a \text{ Dst}(t)$$

- This model has 1 free parameter and gives us a correlation coefficient of 0.978790 on 40 years of data

- Now, we add a second parameter

$$\text{Dst}(t+dt) = a \text{ Dst}(t) + b \text{ VBs}(t)$$

- This model has 2 free parameters and gives us a correlation coefficient of 0.983638

- Finally, the 3 parameter model gives us a correlation coefficient of 0.983645

- The third parameter only added 0.000007 to the correlation coefficient.

– Do we keep the 3rd term, stop, or continue adding new terms?

The Termination Condition

- We stop adding terms when the next term adds so little to the correlation coefficient that it could've occurred by chance
- Specifically, we compare r_n and r_{n+1} to determine whether r_{n+1} exceeds r_n with 95% confidence, given the statistical uncertainty in r_n .

$$\int_{-1}^{r_{n+1}} p(r|\rho = r_n, N) dr \geq 0.95.$$

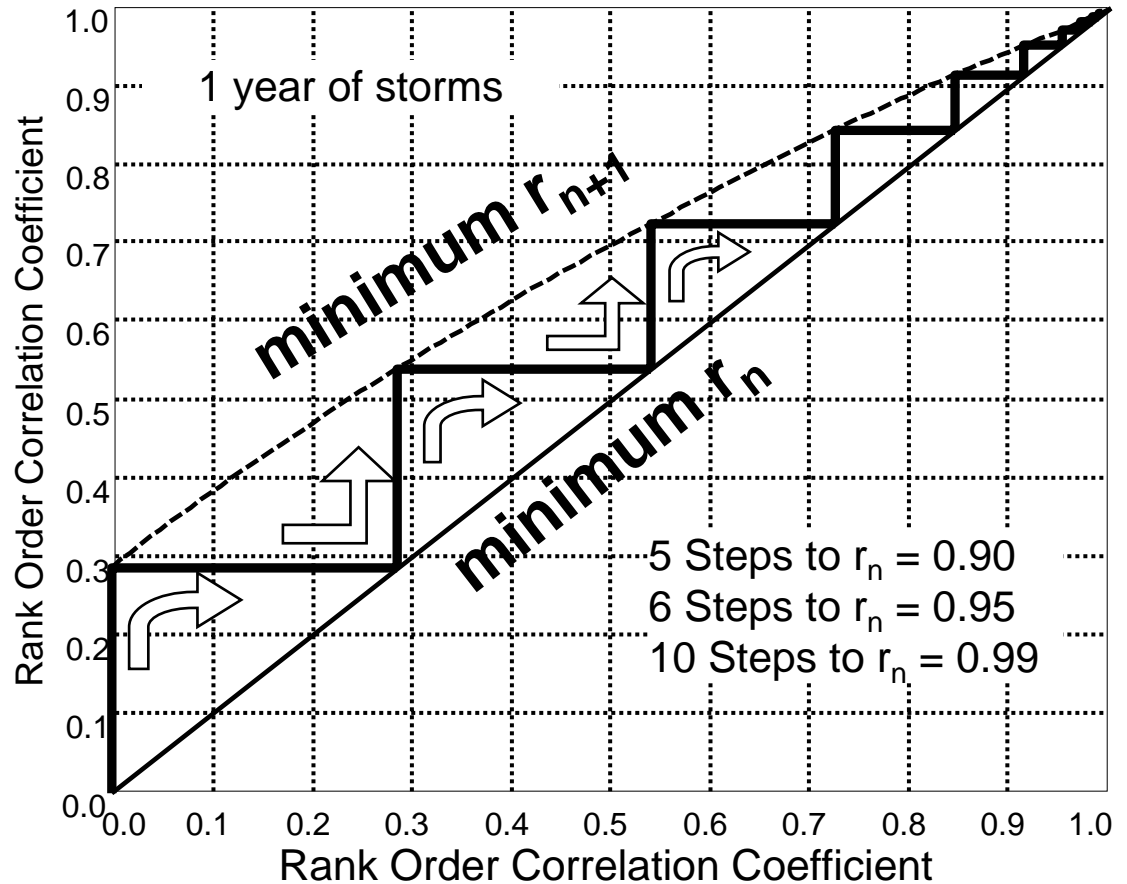
(N is the sample size and ρ is the assumed “true” value of r_n)

- This integral is not easy to evaluate because we don't know $p(r|\rho=r_n, N)$
- But, using reasonable assumptions, we can get a pretty good estimate of $p(r|\rho=r_n, N)$ and construct Occam's Staircase

Occam's Staircase (I)

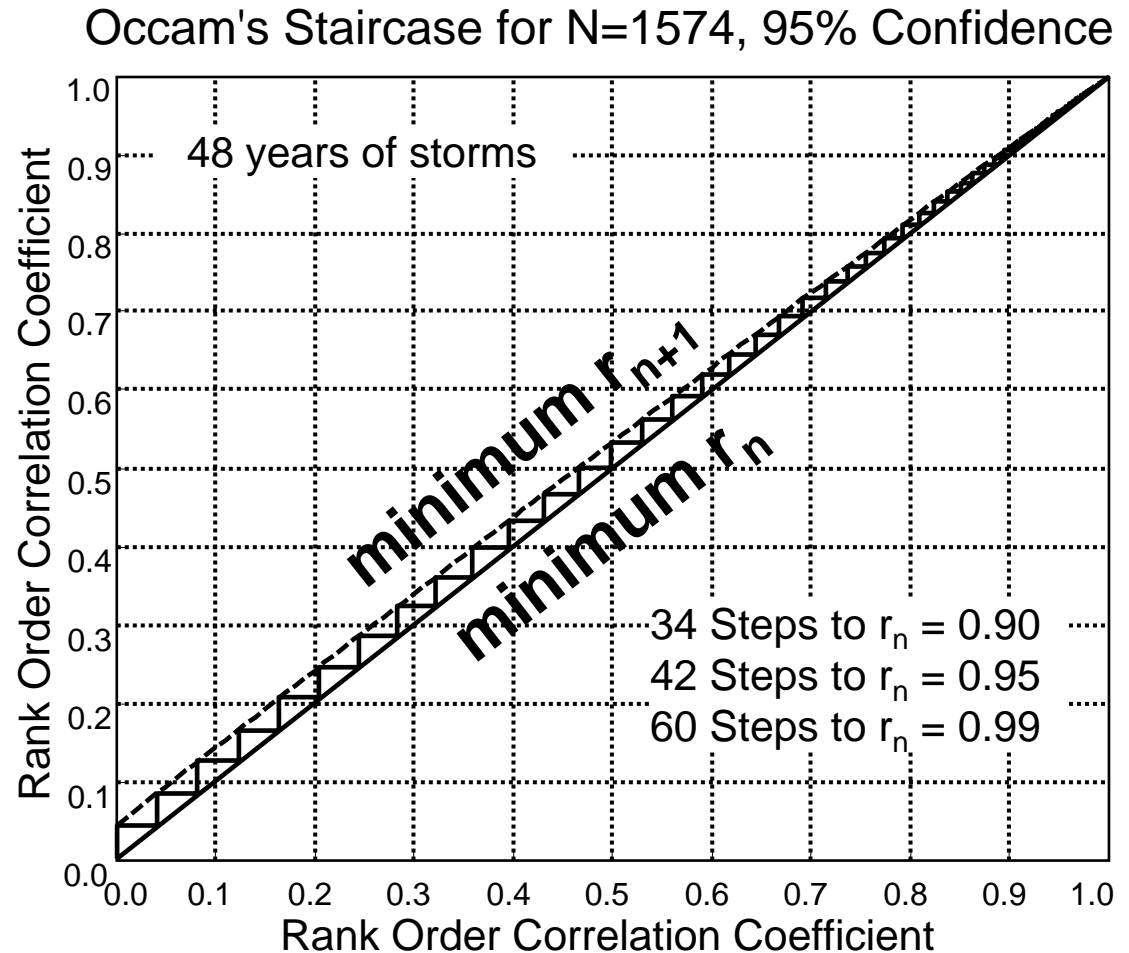
- For a given sample size the iteration from r_n to r_{n+1} creates a staircase progression up a graph
 - e.g., $N=33$, or about 1 year of magnetic storms
- The graph at right depicts the limiting case—the smallest increment r_n to r_{n+1}
 - We start in the lower left corner with 0 free parameters and 0 correlation, $r_0 = 0$.
 - We add one free parameter and move up,
 - Then we move right and use that correlation as the starting point for the addition of the next free parameter
- As n and r^n increases, the step size decreases

Occam's Staircase for $N=33$, 95% Confidence



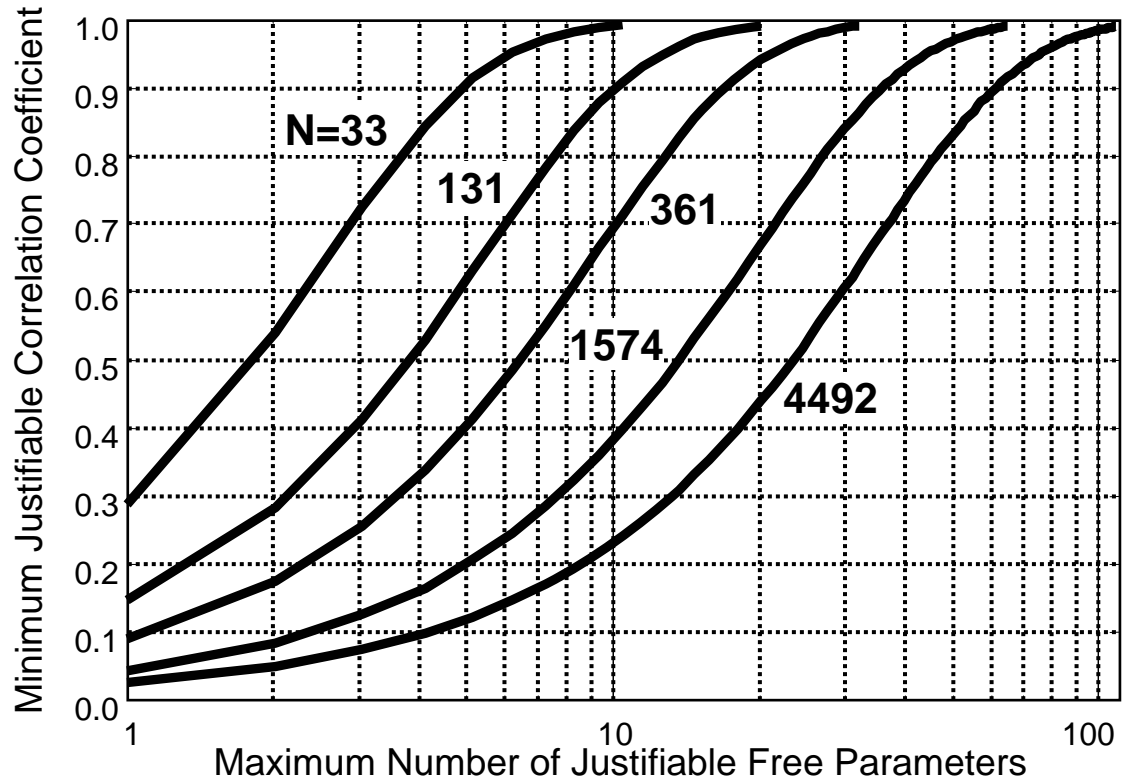
Occam's Staircase (II)

- With $N=1574$, or 48 years of storms...
- The steps are smaller, so we cant take more of them
- More free parameters are justified for a given correlation coefficient



Some Rules of Thumb

- Choosing N for various intervals, we can estimate the limiting relationship between the correlation achieved and the number of free parameters used
- There is roughly a \sqrt{N} dependence in the number of free parameters n and the sample size N for any particular achieved correlation coefficient
- Models with many free parameters must have either
 - Very high correlation coefficient
 - Or, very long data sets



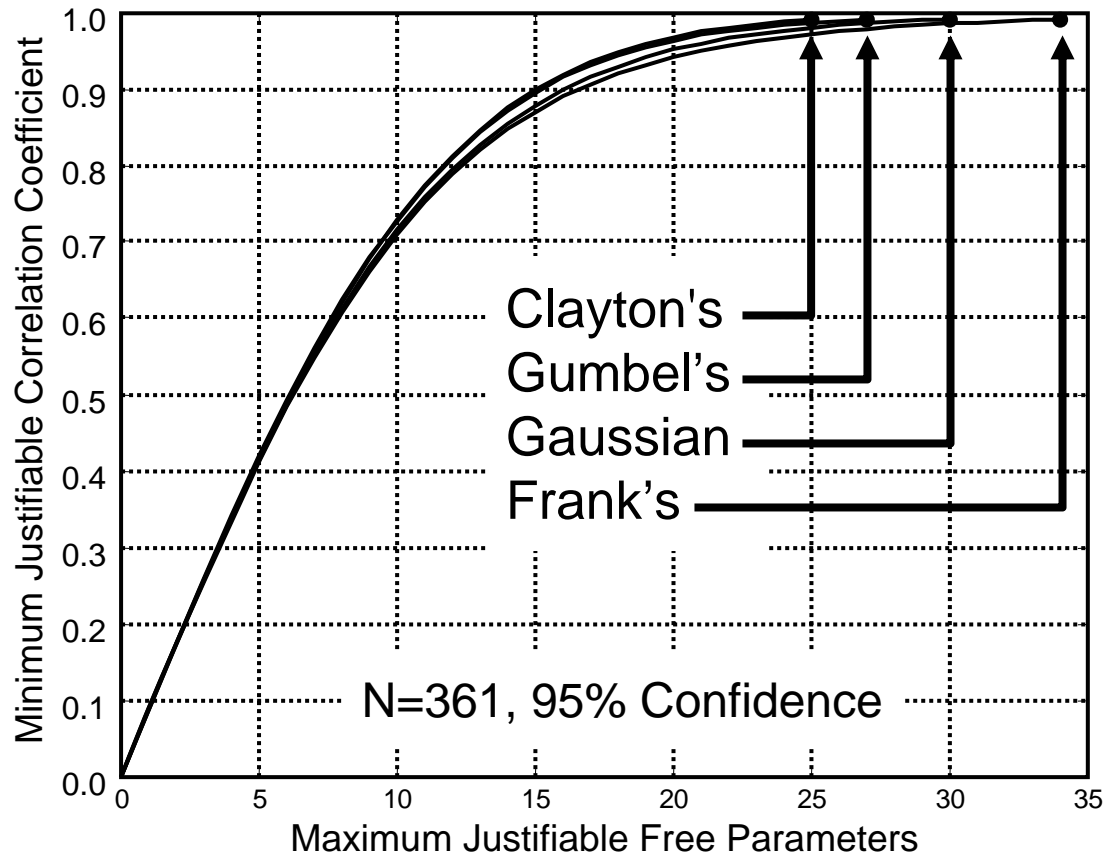
Number of Years	Number of Storms	Steps to ROCC of			Example Observations
		0.90	0.95	0.99	
1	33	5	6	10	CRRES
4	131	10	13	19	IMAGE, POLAR/CAMMICE
11	361	17	20	30	GOES, SAMPEX, WIND
48	1574	34	42	60	<i>Dst</i> index, OMNI database
137	4492	57	70	99	<i>aa</i> index

Conclusion

- **We are often limited more by our data sets than by our empirical modeling methods**
 - **Serial Correlation is a killer**
- **We need:**
 - **Better physical insight to reduce reliance on free parameters**
 - **Longer in situ data sets (e.g., longer NASA missions)**
 - **Inputs that are better correlated with our outputs (Lpp, Dst, electron flux)**
 - E.g., convection electric fields, plasma sheet properties, electron fluxes vs energy and L**

BACKUPS

Occam's Staircase for Four Copulas



Samples from Four Copulas with $\rho = 0.70$

