# New Developments in Time Series Analysis

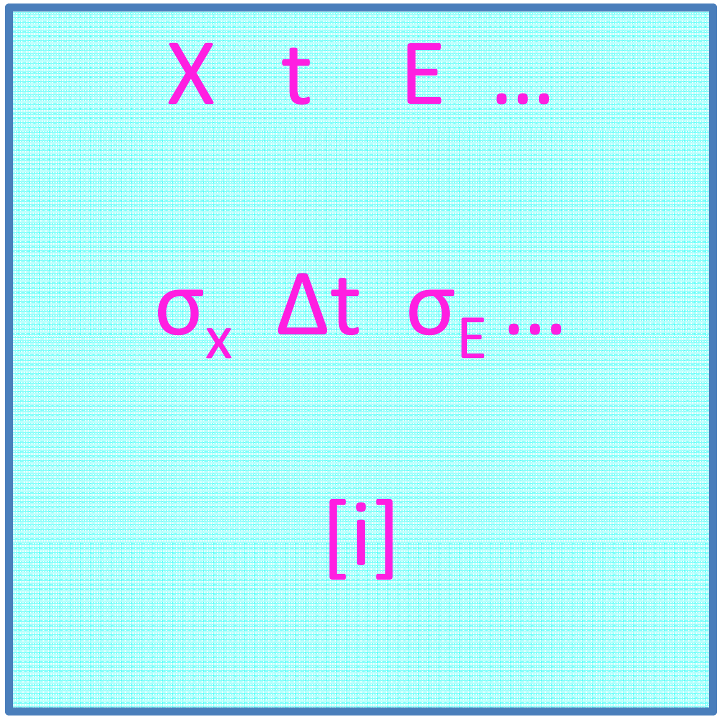
Jeff Scargle

NASA Ames Research Center

- Data Cell Representation
- Time Series Models
- Time-Domain Analysis
- Frequency-Domain Analysis
- Time-Scale/Frequency Distributions

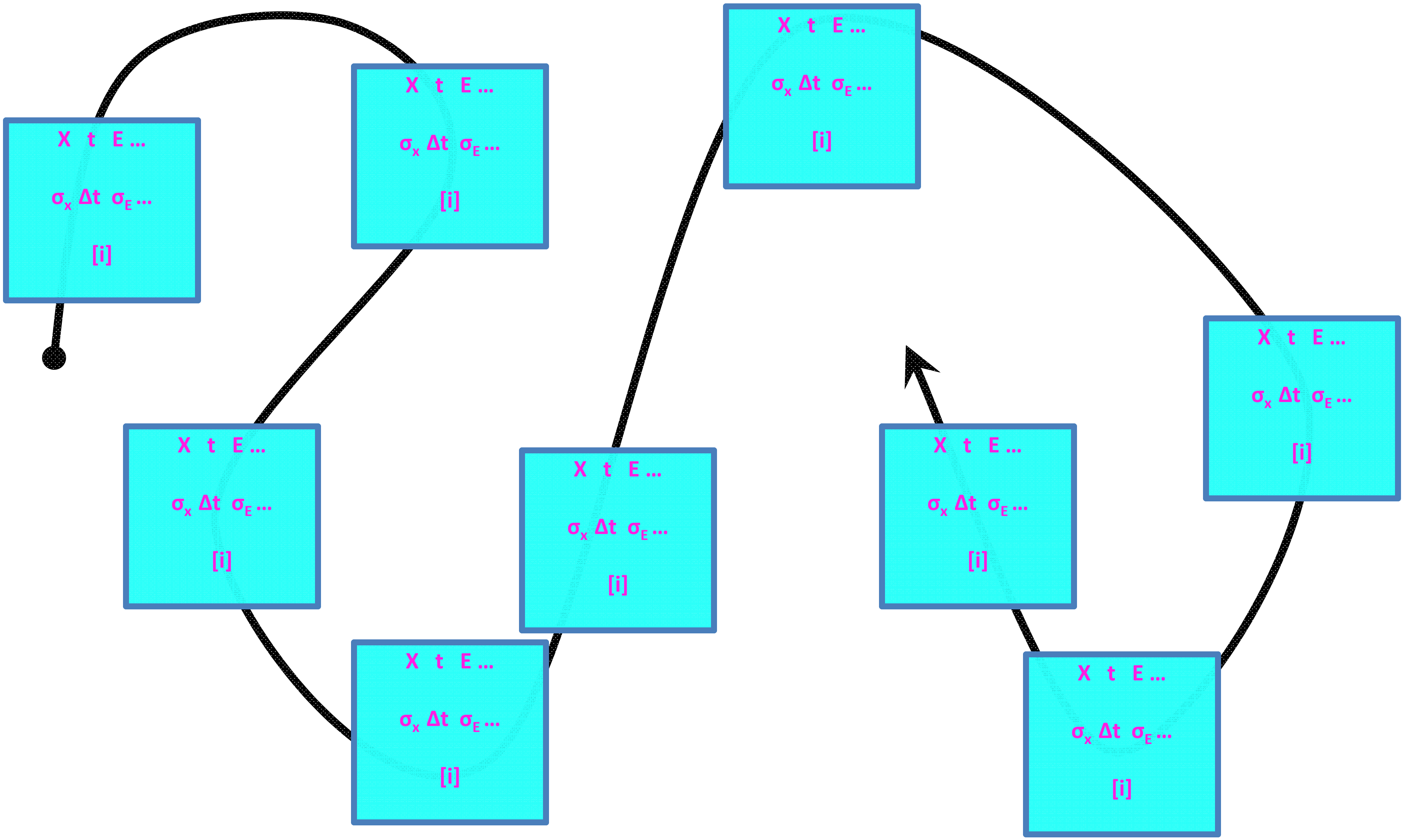
***Digging Deeper: Algorithms for Computationally-Limited Searches in Astronomy***
***June 7, 2011  Keck Institute for Space Studies***

# Data Cell

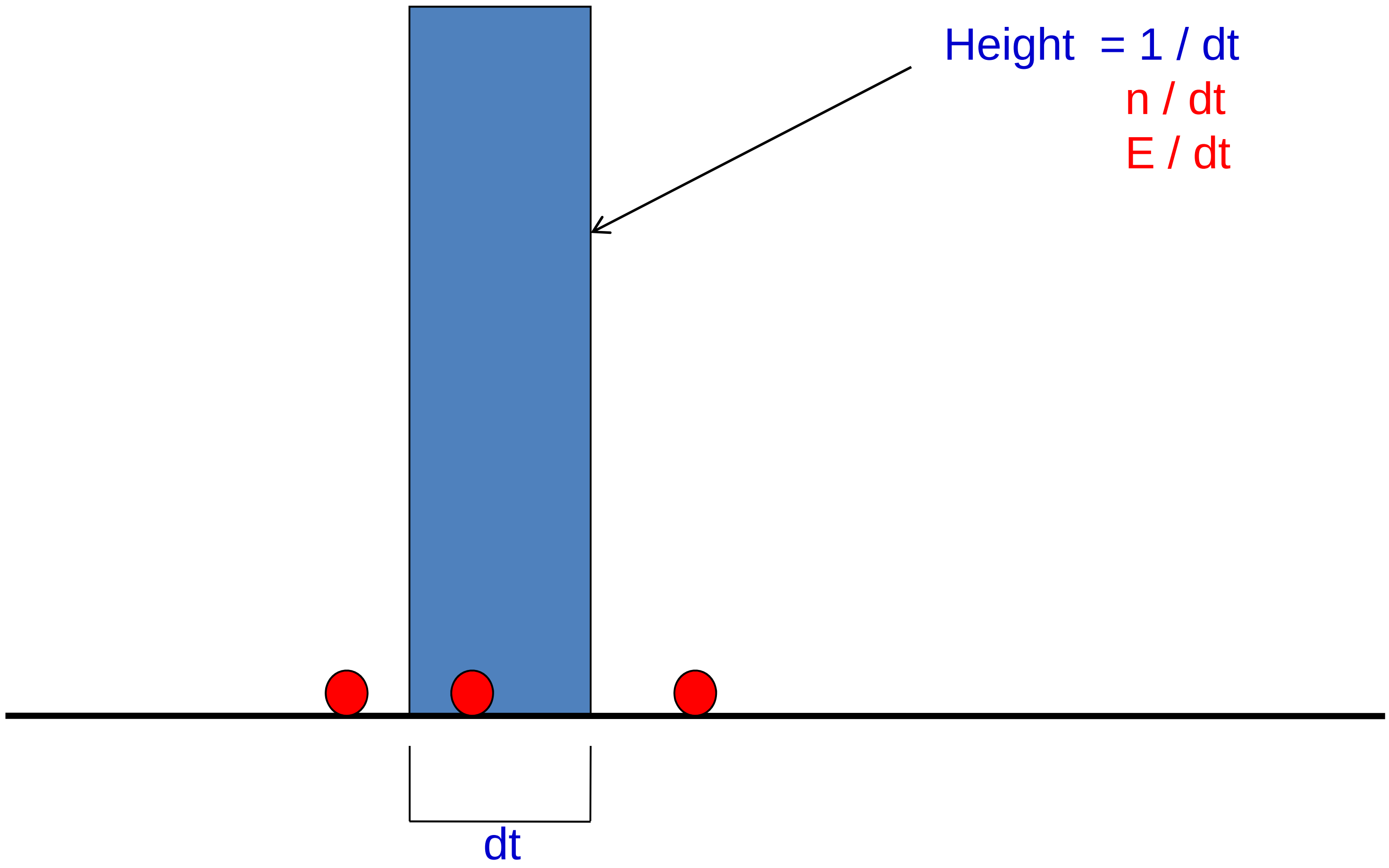



Contains all information relevant to an analysis task.
Index $i = 1, 2, \ldots, N$ denotes individual observations.

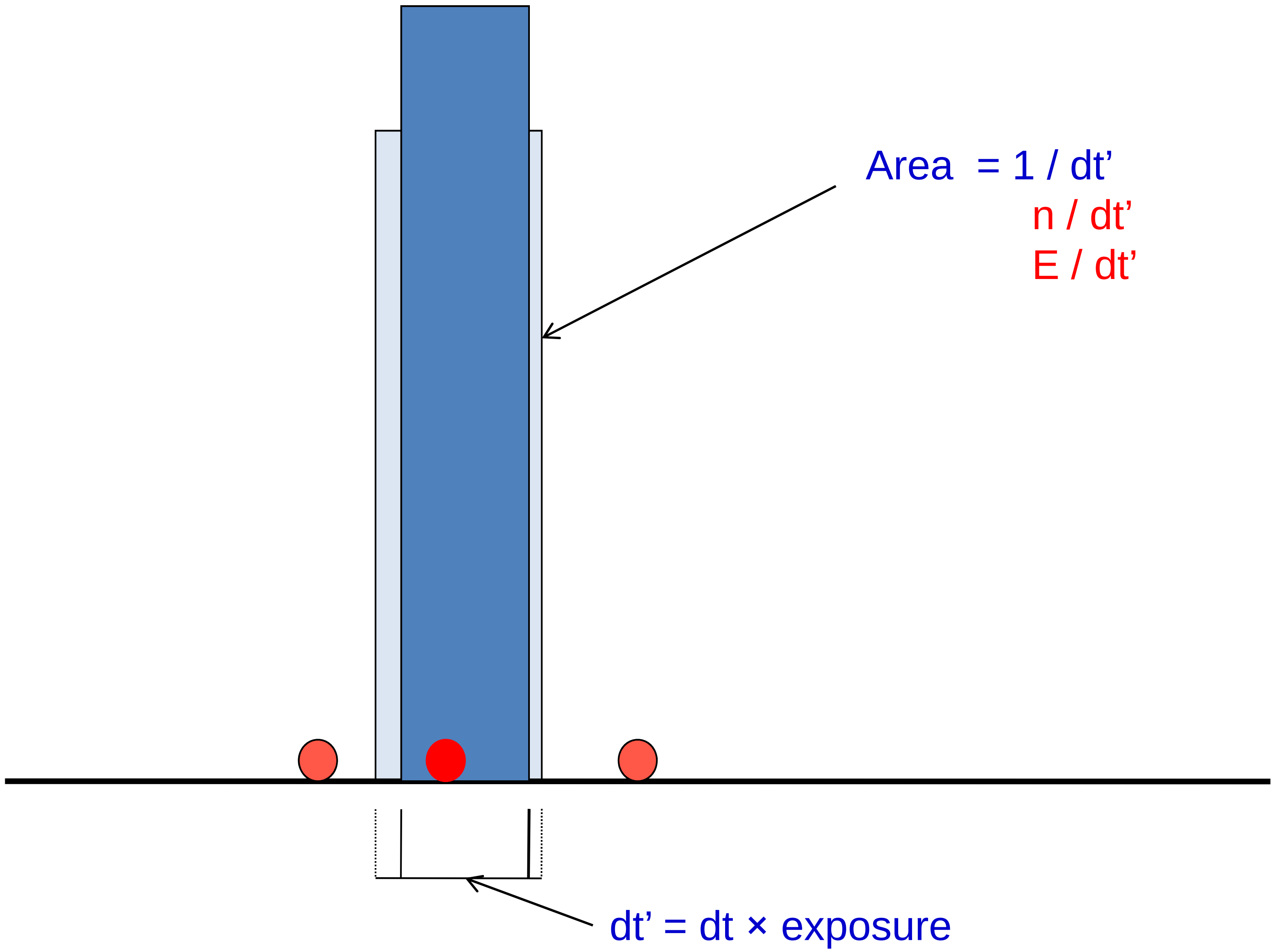
# Time Series: Time-ordered Data Cells




N arbitrary time-ordered data cells make a time series.



Height = 1 / dt
n / dt
E / dt
dt



Area = 1 / dt’
n / dt’
E / dt’
dt’ = dt × exposure

# ***Models***

# The Wold Decomposition Theorem

**Herman Wold (1908–1992) Econometrician**

$X(t) = C * R(t) + D(t)$


Any stationary process is the sum of:

- a moving average (white noise R input to a filter C)
- a linearly deterministic process D

Pulse shape C is …

- Convergent
- Causal, by assumption
- Minimum delay
- Constant

Pulse amplitudes R are …

- Uncorrelated ("white")

An extension of the theorem (Scargle 1981) relaxes some restrictions

***Powerful existence proof of a random flare representation for any stationary time series***

# Testing for Stationarity

- Formal definition requires infinite amount of data
- Local stationarity depends on scale
- Construct stationarity measure S[ x(t) ]
  - E.g. variance of TF distribution vs. time marginal
  - Any such measure has statistical fluctuations
  - Simulate surrogate data: scramble Fourier phase
- Construct distribution of S( surrogate data )

*Testing Stationarity with Time-Frequency Surrogates*, Jun Xiao, Pierre Borgnat, and Patrick Flandrin

From:
Flandrin & Borgnat
"Revisiting and testing stationarity," 2008

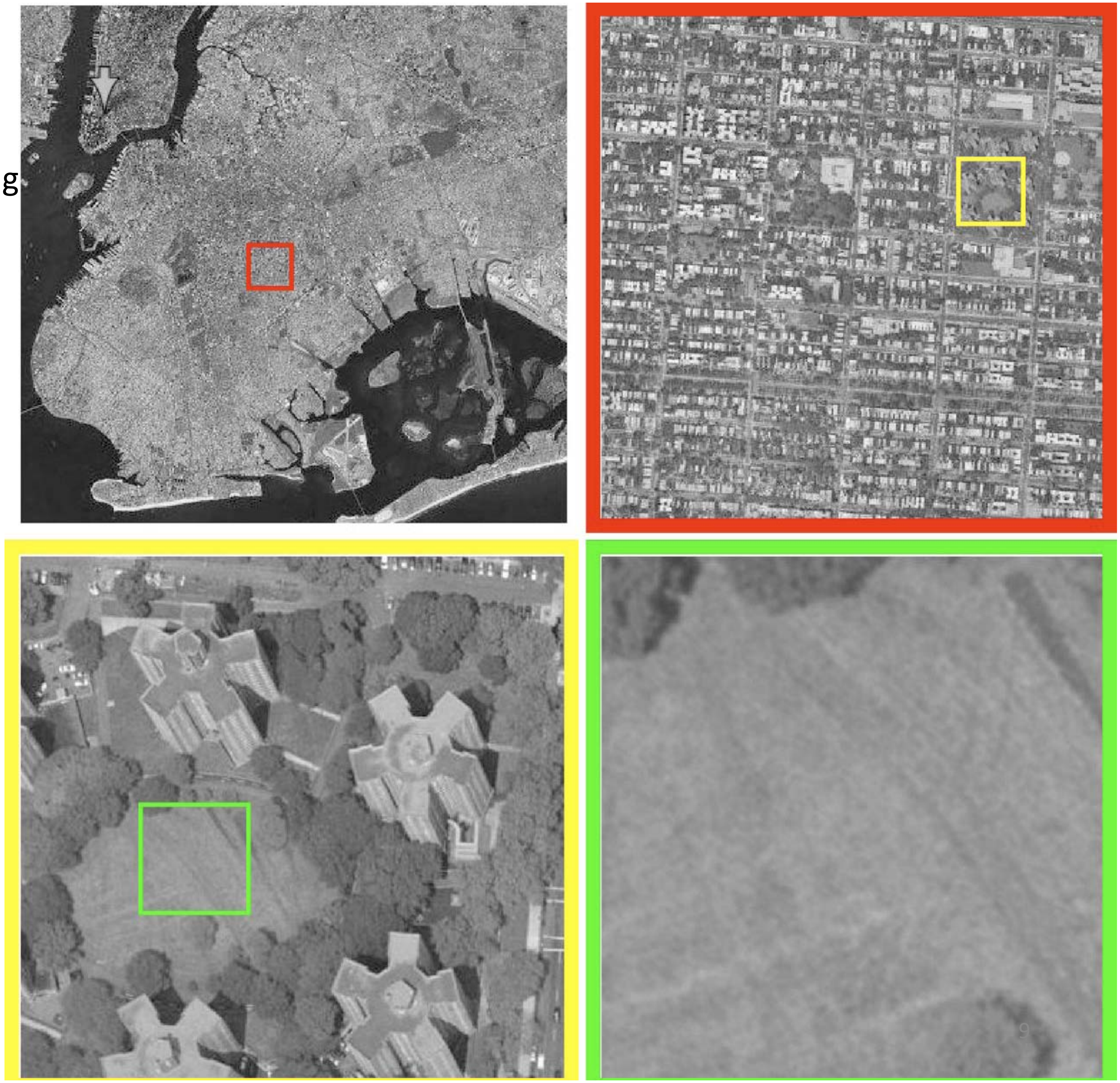
… interpreted as "stationary" or "nonstationary" depending on the observation scale …

TL: nonstationary

TR: stationary (periodic)

BL: nonstationary

BR: stationary (homogeneous texture)



# Processes Involved in Astronomical Time Series

Variable Source → Propagation To Observer → Photon Detection

- Luminosity fluctuations ( Random or Deterministic )
- Photon Emission ( Poisson Process )

- Scintillation
- Dispersion
- …

( Random )

- Photon Detection: ( Poisson Process )

Correlation here … does not mean … Correlation here!

# ***Time Domain***

# Bayesian (or Maximum Likelihood) Blocks

Signal represented as *constant in elements of data-space partition*
Optimize by maximizing model fitness over all possible partitions.

- Nonparametric

- Assumptions: prior on amplitudes and number of blocks

- No limitation on resolution in the independent variable (no bins!)

- *Real-time mode: first significant signal above background*

- Representation, discontinuous, convenient for further analysis

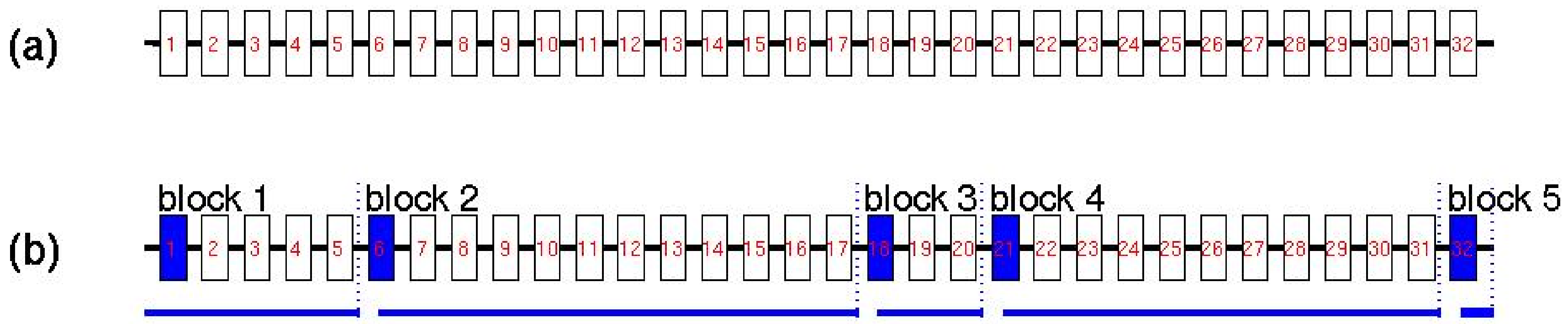
- Local structure, not global

- O $N^2$ ( maybe can be faster )

# Blocks

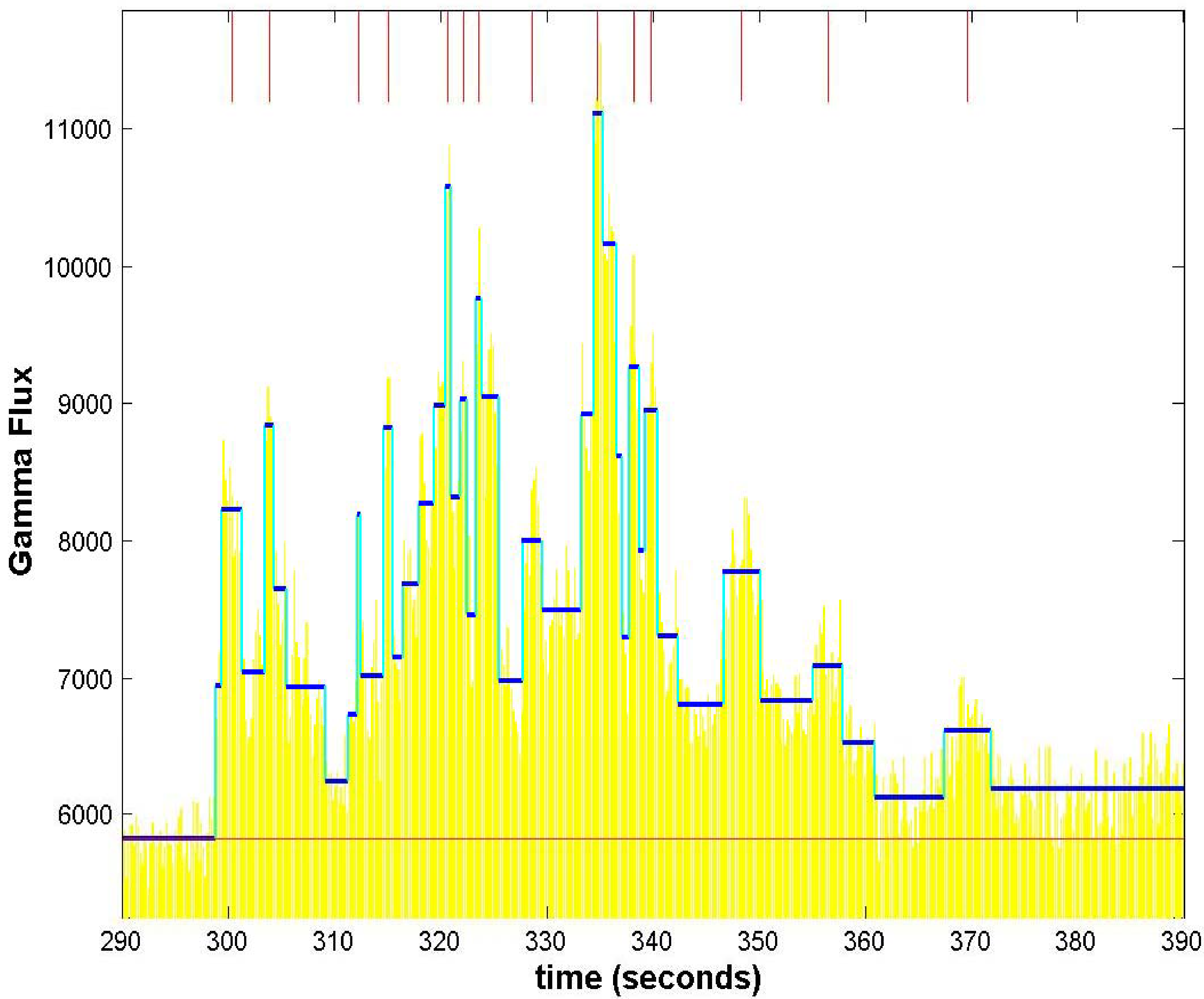
Block: a set of data cells

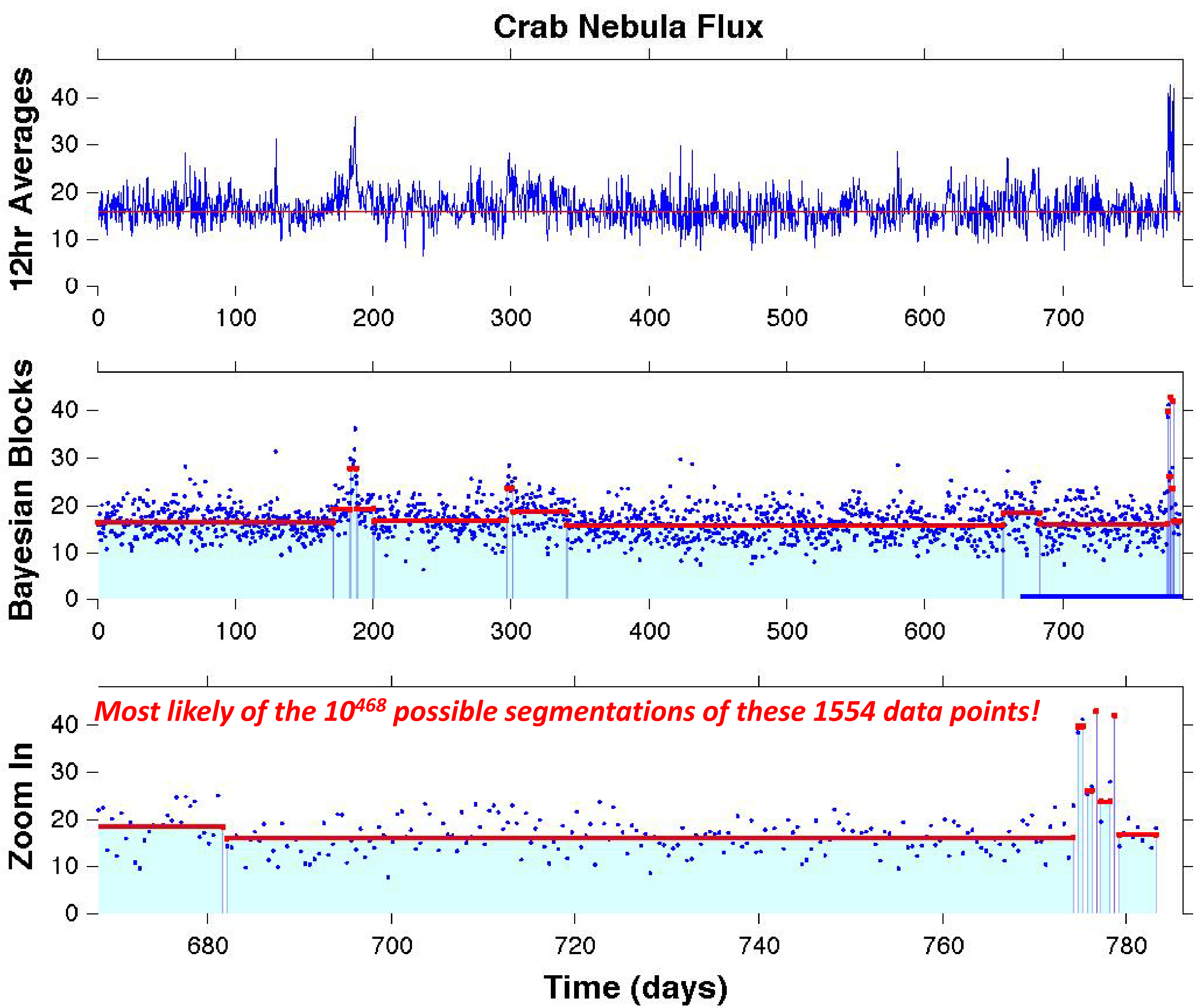
Fitness function = sum over blocks F( Block )



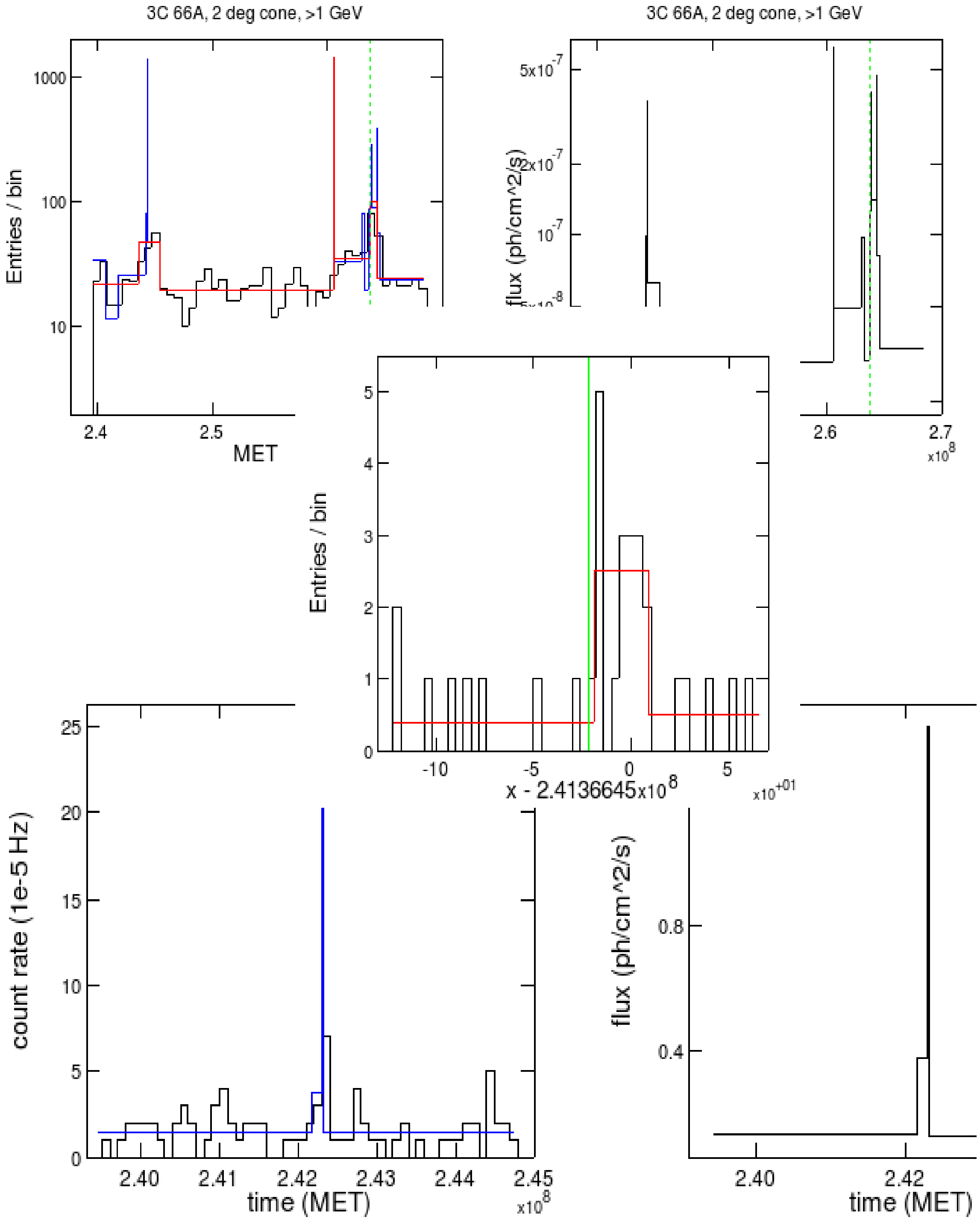
# The Optimizer

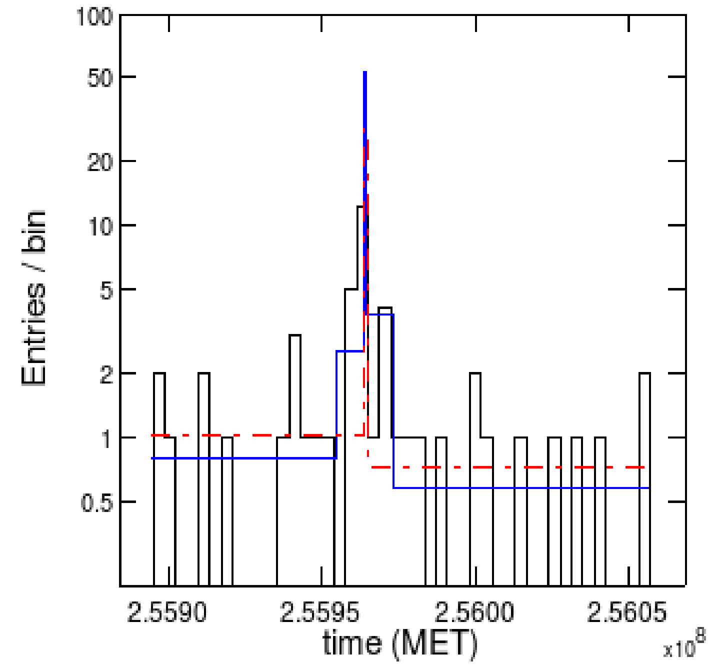
```
best = []; last = [];
for R = 1:num_cells
   [ best(R), last(R) ] = max( [0 best] + ...
          reverse( log_post( cumsum( data_cells(R:-1:1, :) ), prior,  type )  ) );

    if first > 0 & last(R) > first  % Option: trigger on first significant block
          changepoints = last(R); return
    end
end

% Now locate all the changepoints
index = last( num_cells );
changepoints = [];
while index > 1
   changepoints = [ index changepoints ];
    index = last( index - 1 );
end
```

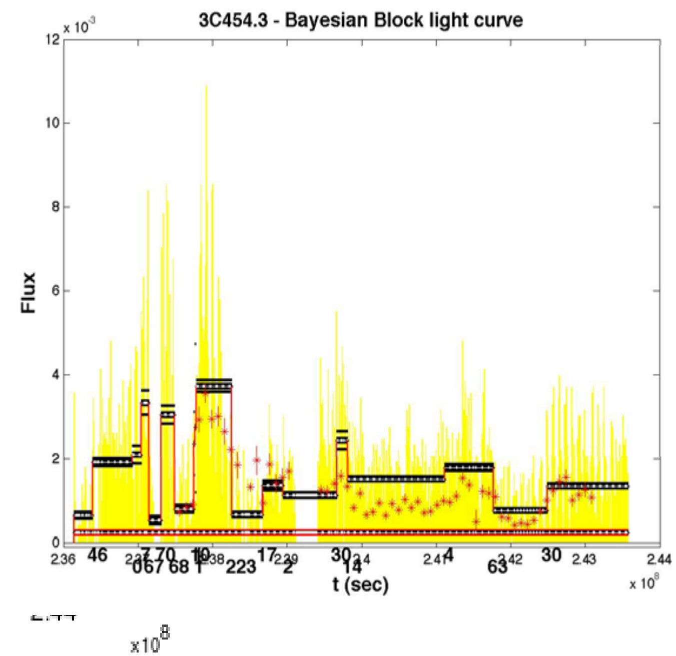

Gamma Flux
11000
10000
9000
8000
7000
6000
290
300
310
320
330
340
350
360
370
380
390
time (seconds)

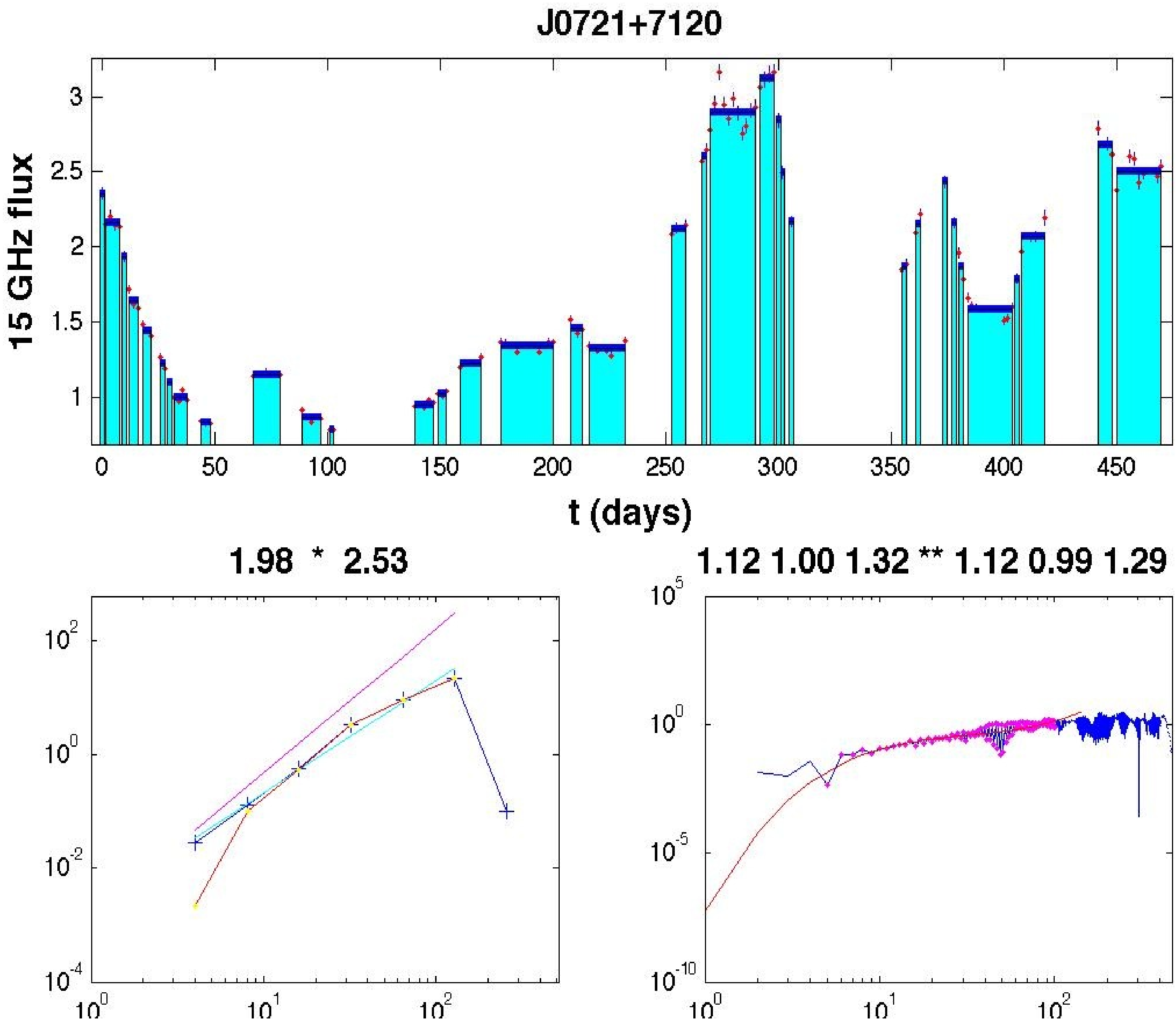


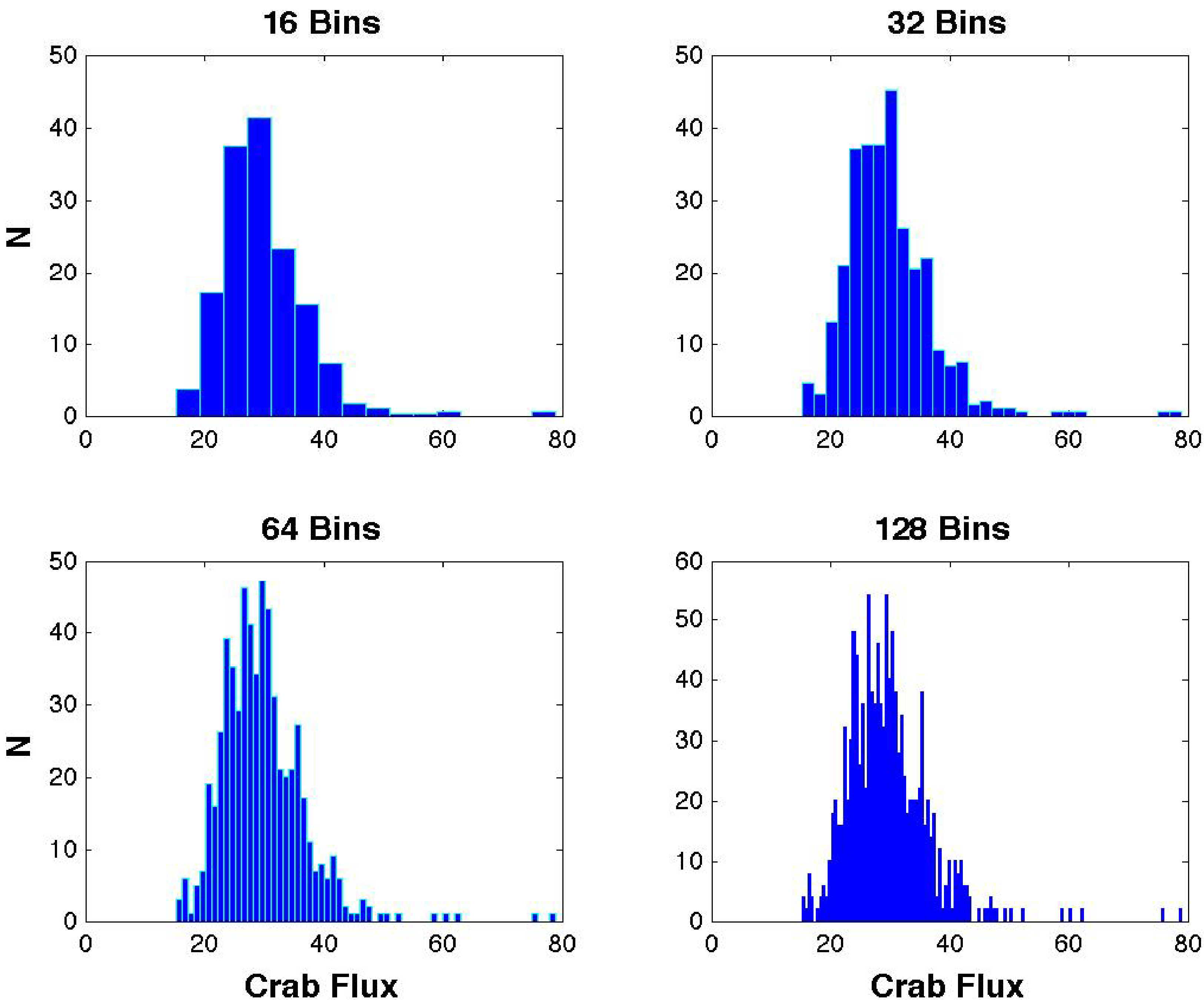
Crab Nebula Flux
12hr Averages
Bayesian Blocks
Zoom In
Most likely of the 10^468 possible segmentations of these 1554 data points!
Time (days)

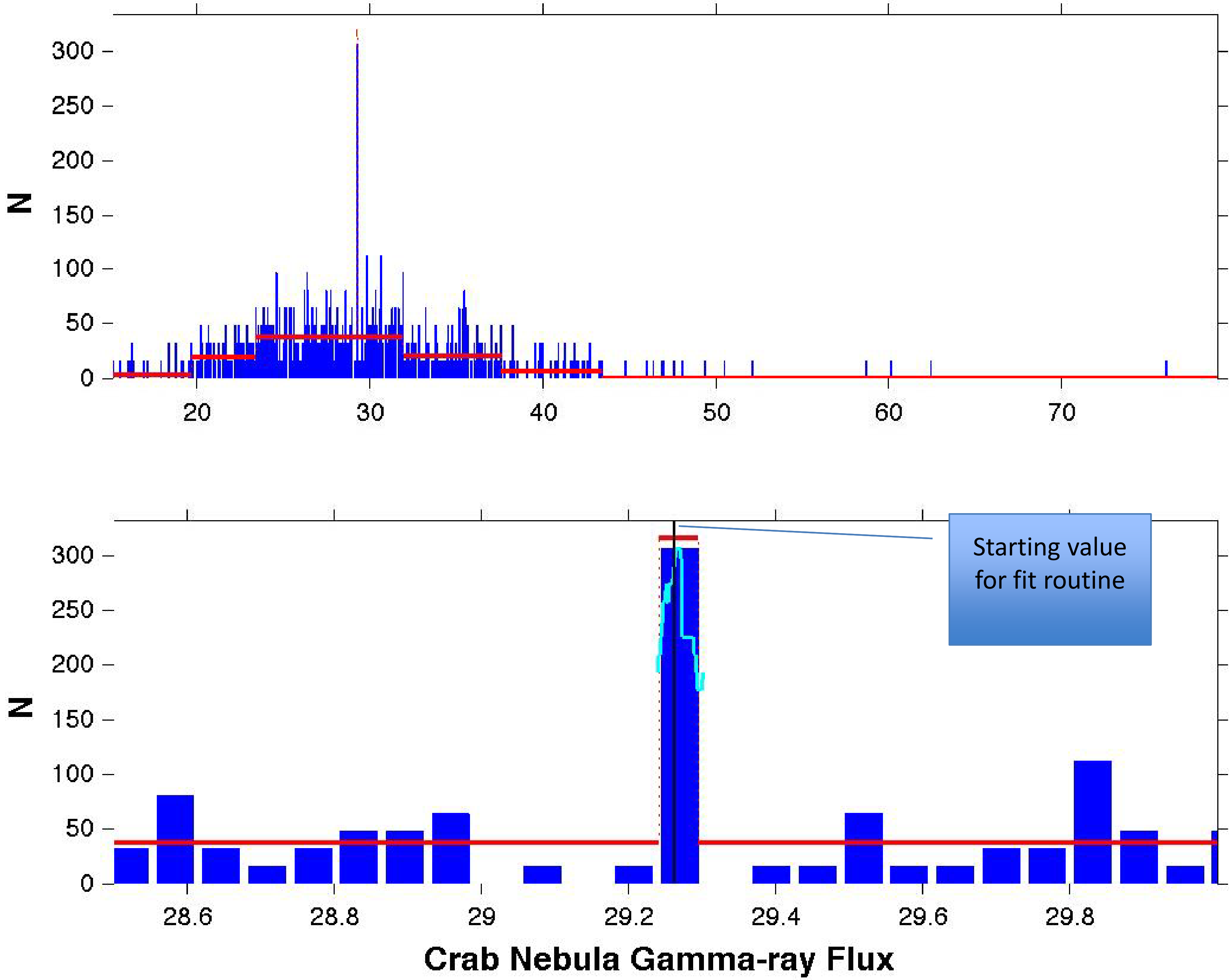

3C 66A, 2 deg cone, >1 GeV
3C 66A, 2 deg cone, >1 GeV
Entries / bin
MET
flux (ph/cm^2/s)
x - 2.4136645x10^8
count rate (1e-5 Hz)
time (MET)
flux (ph/cm^2/s)
time (MET)


Entries / bin
time (MET)


3C454.3 - Bayesian Block light curve
Flux
t (sec)


J0721+7120
15 GHz flux
t (days)
1.98 * 2.53
1.12 1.00 1.32 ** 1.12 0.99 1.29


16 Bins
32 Bins
64 Bins
128 Bins
N
N
Crab Flux
Crab Flux
0
20
40
60
80
10
30
50



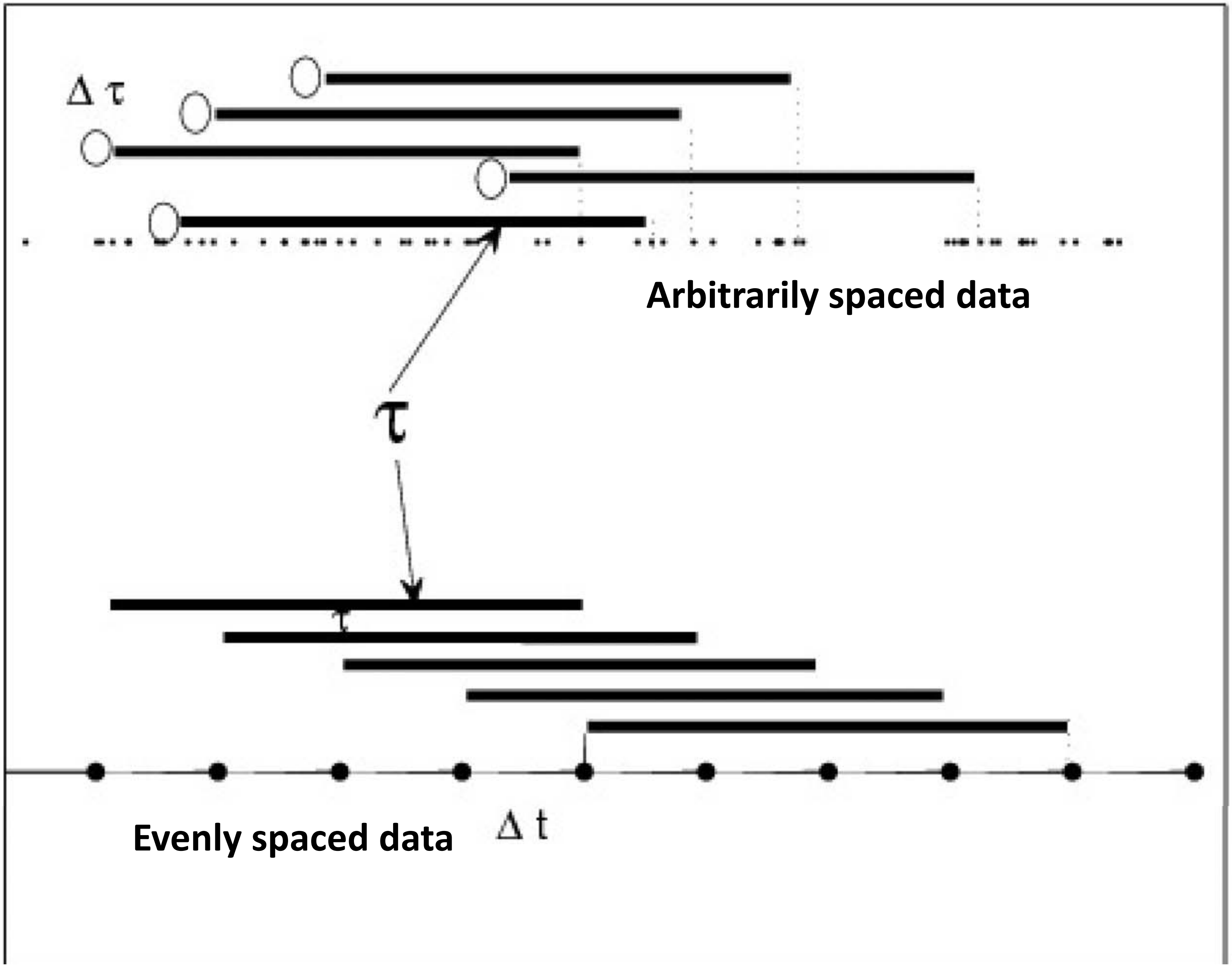
N
300
250
200
150
100
50
0
20
30
40
50
60
70
N
300
250
200
150
100
50
0
28.6
28.8
29
29.2
29.4
29.6
29.8
Starting value for fit routine
Crab Nebula Gamma-ray Flux

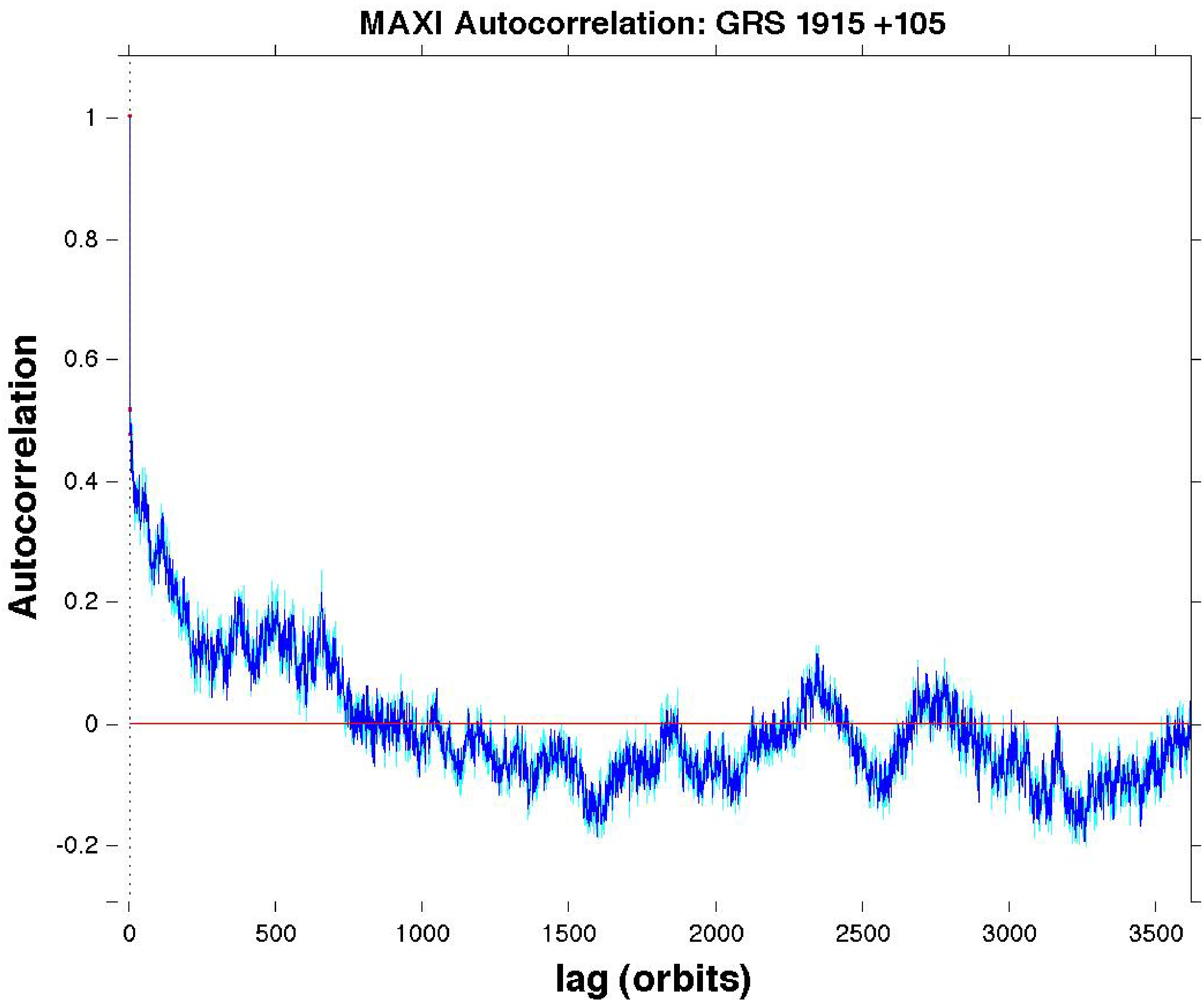
# Cross- and Auto- Correlation Functions for unevenly spaced data

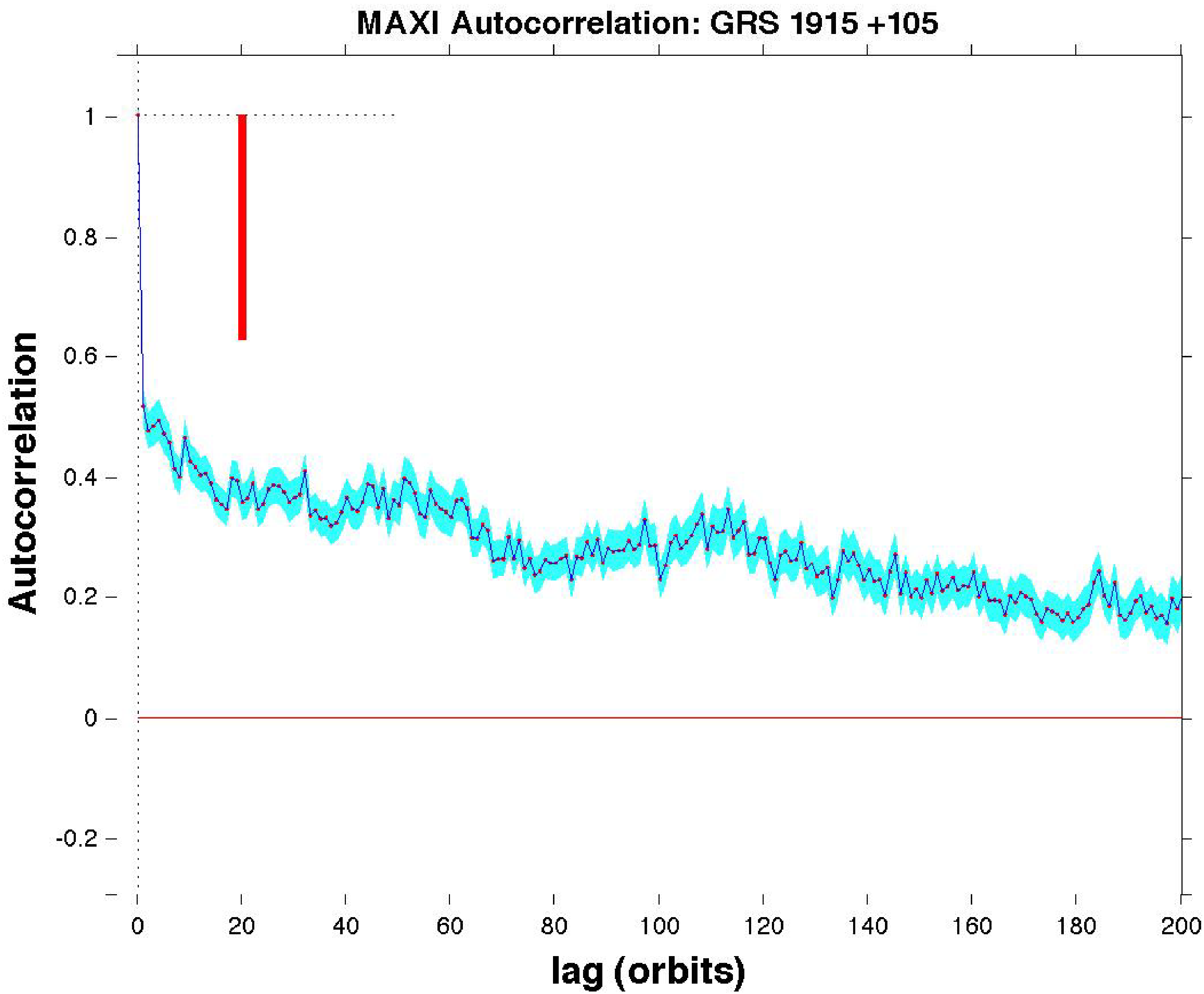
## Edelson and Krolik:

The Discrete Correlation Function: a New Method for Analyzing Unevenly Sampled Variability Data
Ap. J. 333 (1988) 646

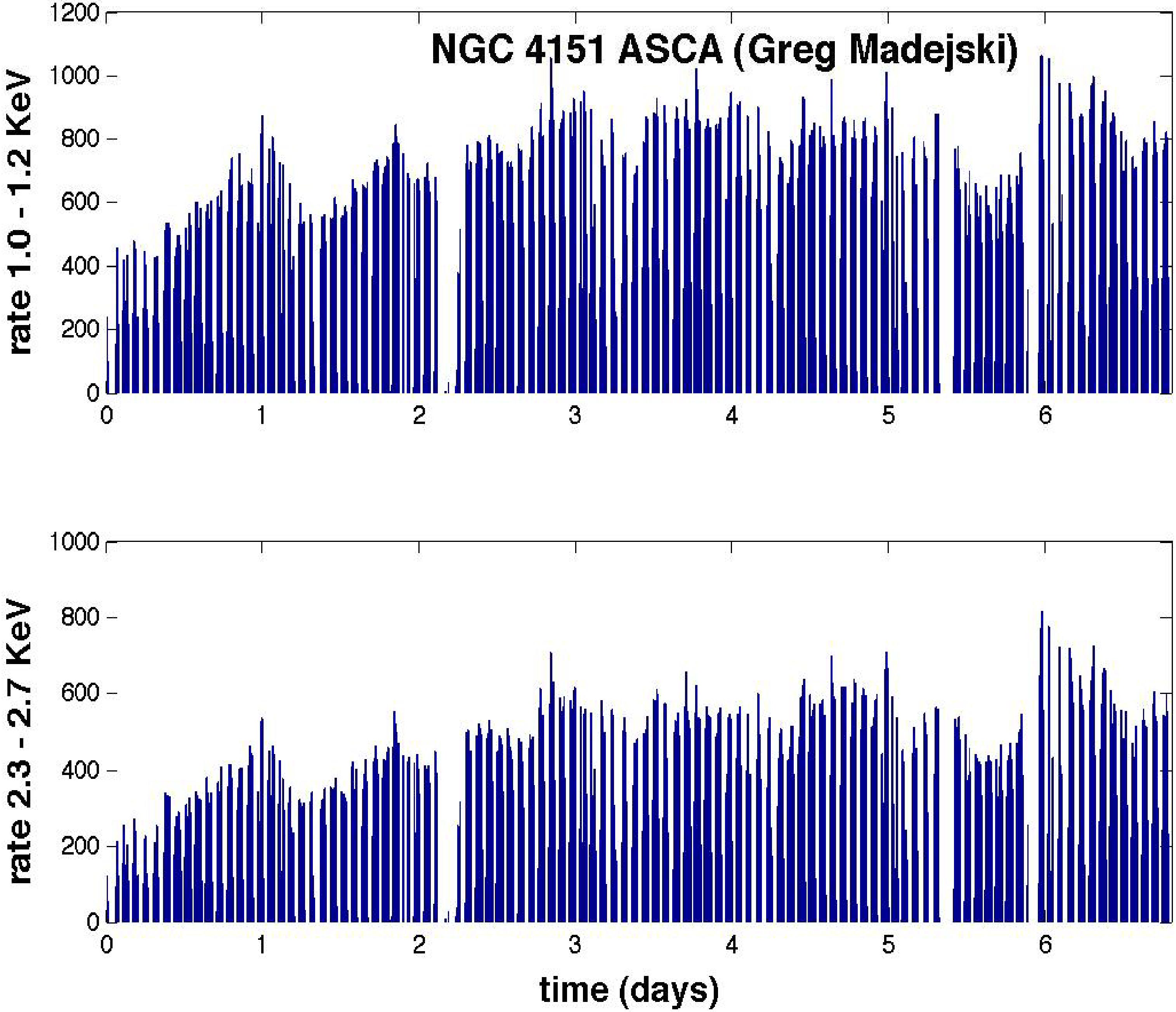
Edelson and Krolik: The Discrete Correlation Function: a New Method for Analyzing Unevenly Sampled Variability Data, Ap. J. 333, 1988, 646 - starting point for all else!

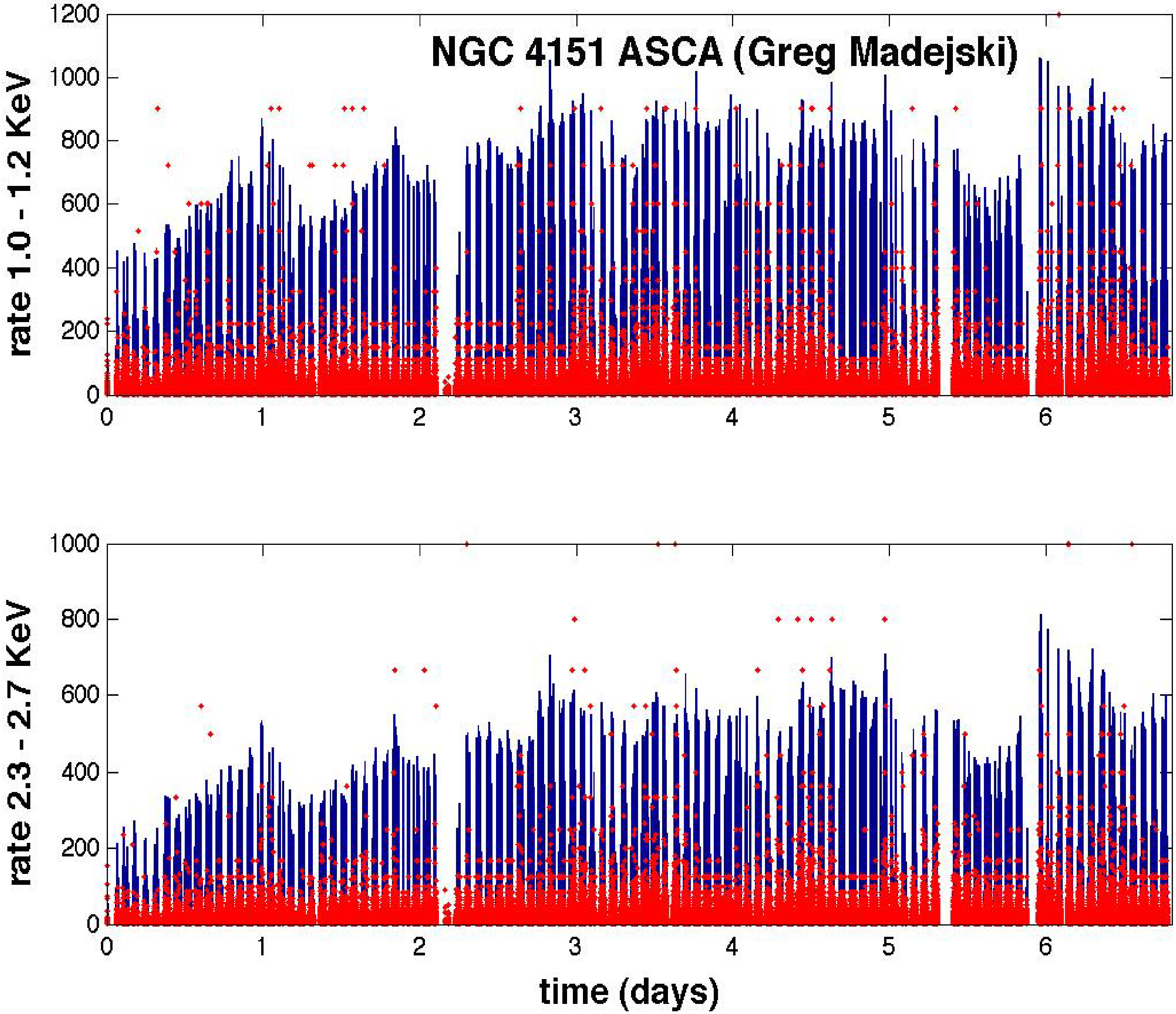


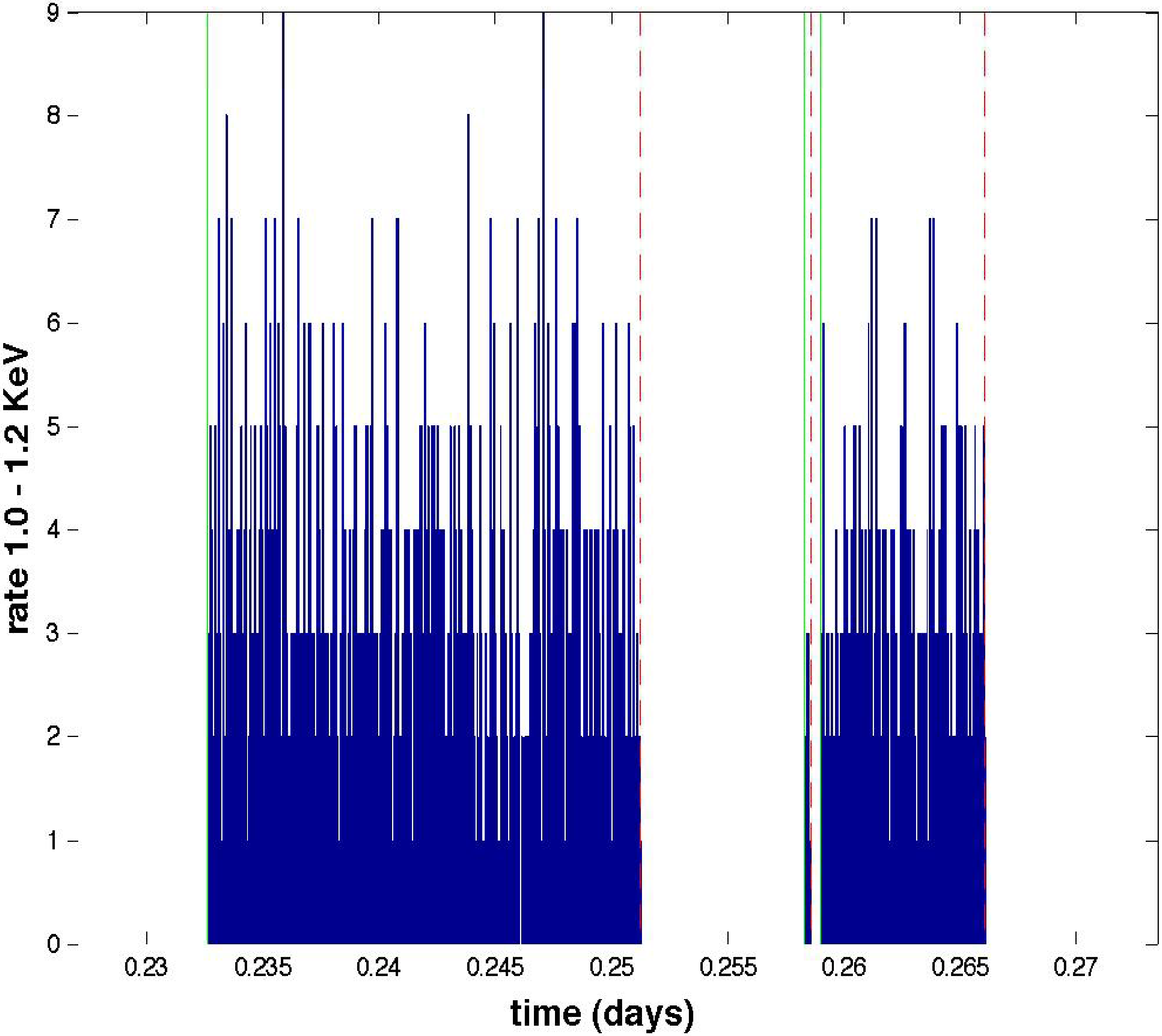

MAXI Autocorrelation: GRS 1915 +105
Autocorrelation
lag (orbits)
1
0.8
0.6
0.4
0.2
0
-0.2
0
500
1000
1500
2000
2500
3000
3500

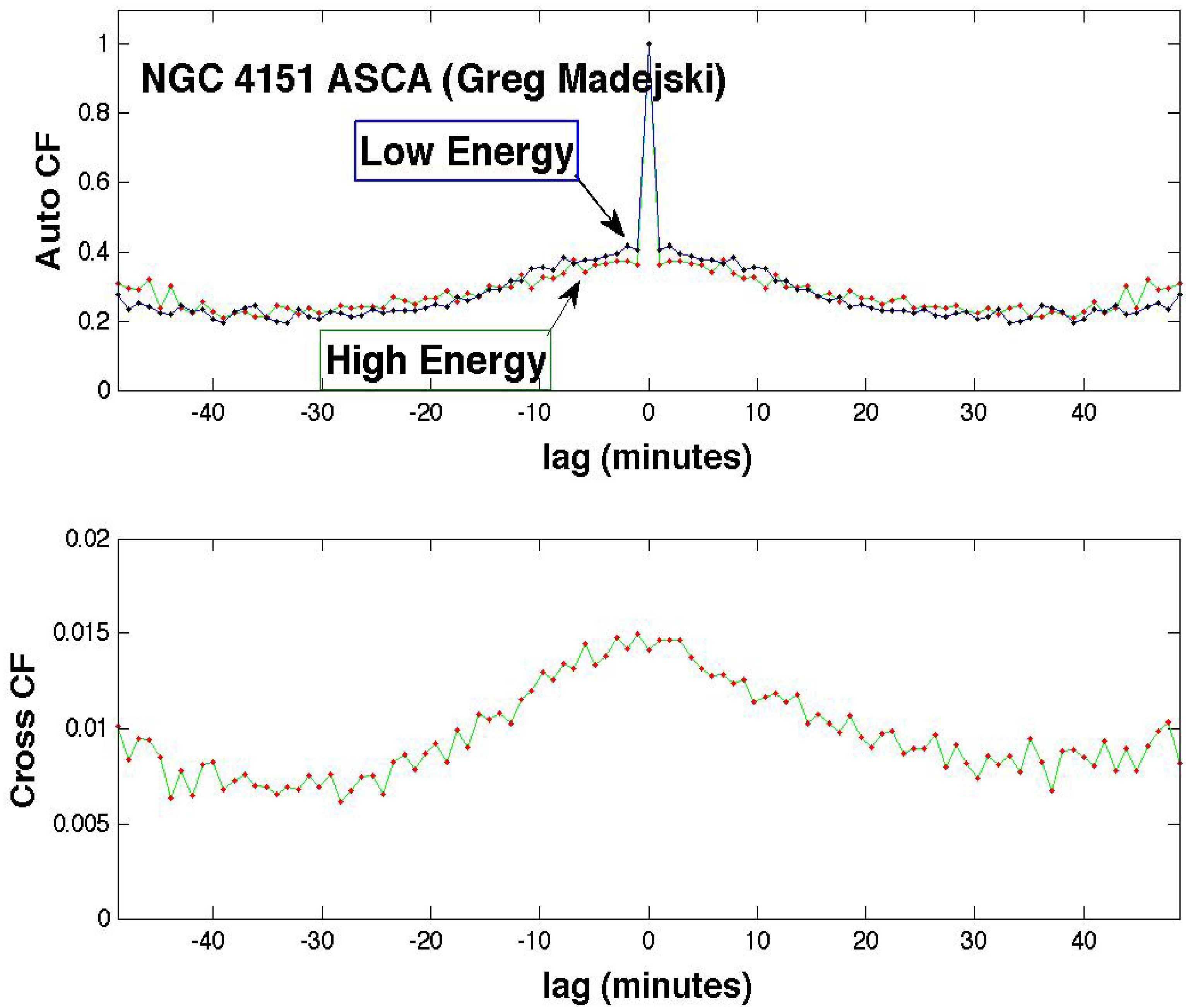


MAXI Autocorrelation: GRS 1915 +105
Autocorrelation
lag (orbits)
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160
180
200


NGC 4151 ASCA (Greg Madejski)
rate 1.0 - 1.2 KeV
1200
1000
800
600
400
200
0
0
1
2
3
4
5
6
rate 2.3 - 2.7 KeV
1000
800
600
400
200
0
0
1
2
3
4
5
6
time (days)


NGC 4151 ASCA (Greg Madejski)
rate 1.0 - 1.2 KeV
rate 2.3 - 2.7 KeV
time (days)
1200
1000
800
600
400
200
0
1
2
3
4
5
6


rate 1.0 - 1.2 KeV
time (days)
0
1
2
3
4
5
6
7
8
9
0.23
0.235
0.24
0.245
0.25
0.255
0.26
0.265
0.27


NGC 4151 ASCA (Greg Madejski)
Low Energy
High Energy
Auto CF
lag (minutes)
Cross CF
lag (minutes)

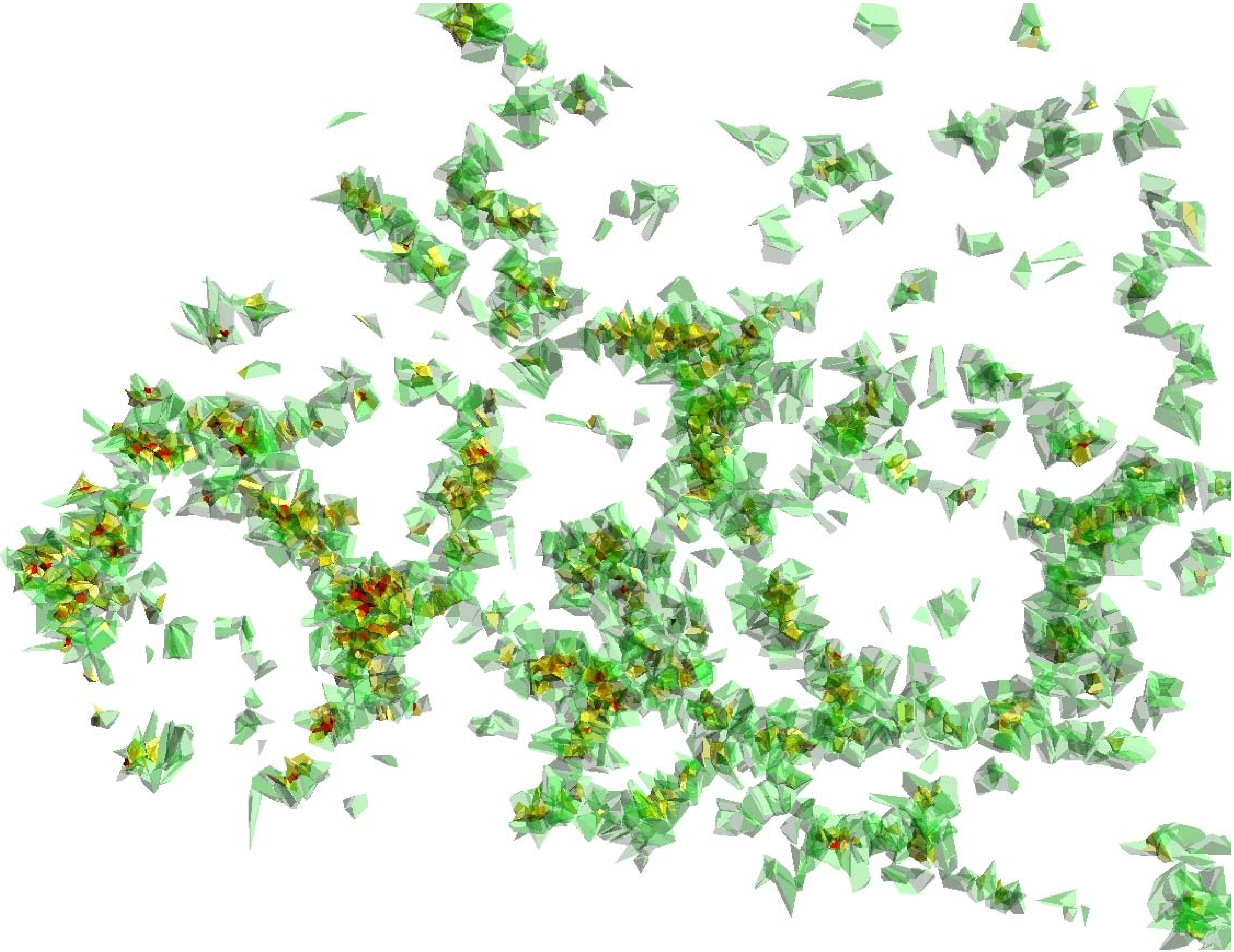
# Event Data Cells in any dimension

| | |
|---|---|
| Measurements: | Point coordinates |
| Data Space: | Space of any dimension |
| Signal: | Point density (deterministic or probabilistic) |
| Data Cell: | Voronoi cells for the data points |

Suf. Statistics N = number of points in block
V = volume of block

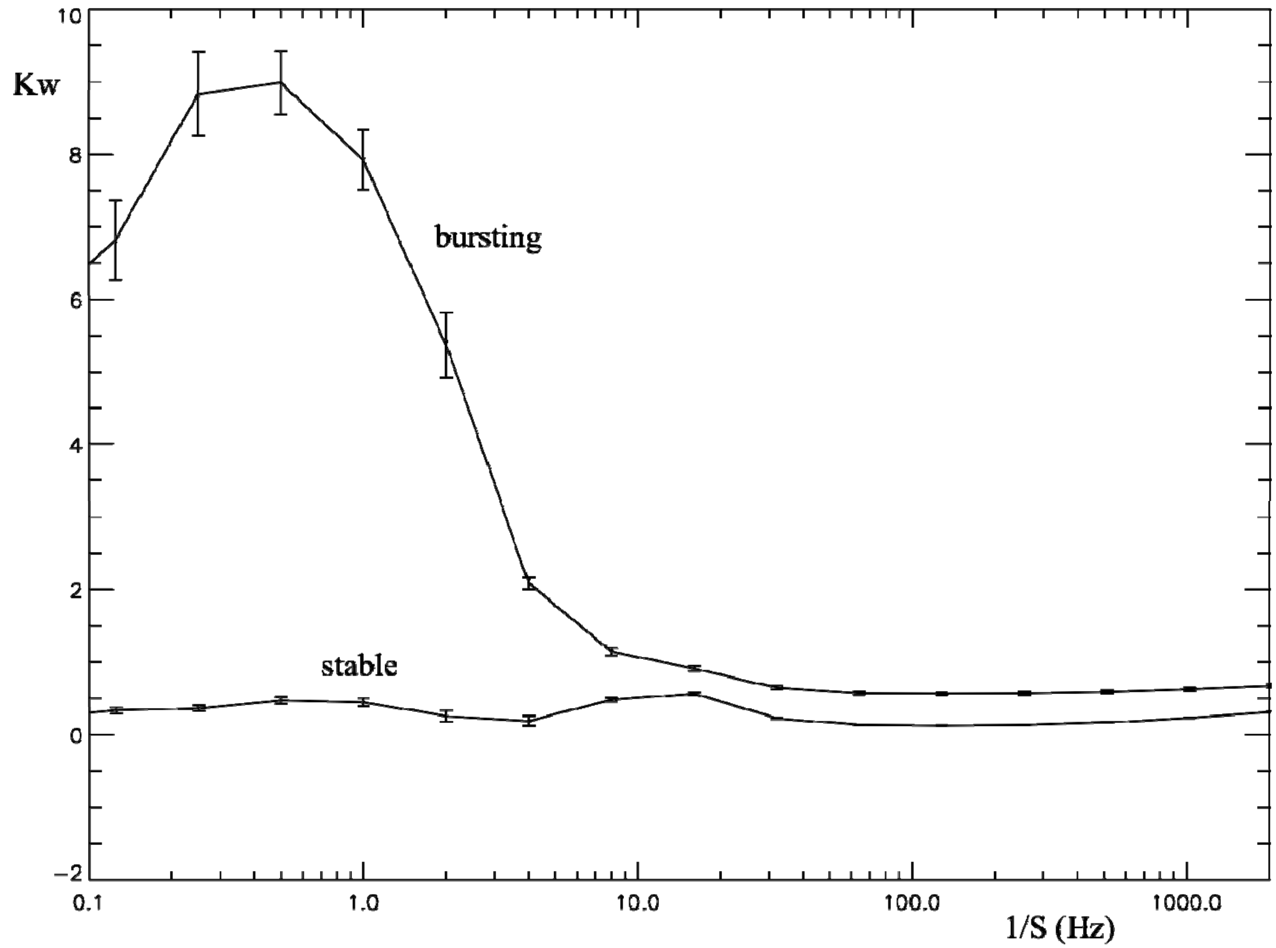
| | |
|---|---|
| Max Likelihood: | $( N / V )^{N} e^{-N}$ |
| Posterior: | $N! (V - N)! / (V+1)!$ |

Example: any problem usually approached with histograms (1D)
positions of objects from a sky survey (2D)
positions of objects in a redshift survey (3D)



# Wavelet Kurtosis

## New Statistic to Detect and Characterize Intermittency




**Daniel Engavatov, Elliott Bloom, JS; SLAC PhD Thesis**

***Frequency Domain***

Data Mode

- Photon events
- Time-to-Spill
- Counts in bins
- Flux measurements
- Any Mode/Sampling!

# Universal Time Series Analysis Machine

Auto-

- Correlation Function
- Fourier Power Spectrum
- Fourier Phase Spectrum
- Wavelet scalgram
- Wavelet scaleogram
- Structure Function
- Time-Frequency Distribution
- Time-Scale Distribution
- …

Extension of Edelson & Krolik
Algorithm for Correlation Function
of Unevenly Sampled Data

Jeff Scargle

Data Mode

- Photon events
- Time-to-Spill
- Counts in bins
- Flux measurements
- Any Mode/Sampling!

***any combination***

- Photon events
- Time-to-Spill
- Counts in bins
- Flux measurements
- Any Mode/Sampling!

# Universal Time Series Analysis Machine

Extension of Edelson & Krolik Algorithm for Correlation Function of Unevenly Sampled Data

## Cross-

- Correlation Function
- Fourier Power Spectrum
- Fourier Phase Spectrum
- Wavelet scalgram
- Wavelet scaleogram
- Structure Function
- Time-Frequency Distribution
- Time-Scale Distribution
- …

Jeff Scargle

***Time-Scale***
***Time-Frequency***

# Time-Frequency/Time-Scale Analysis

***Transform to a new view of the time series information.***

- A Reality in joint time & frequency (or scale) representation
- Atomic decomposition
  - Time-frequency atoms
  - Over-complete representations
  - Optimal Basis Pursuit (Mallat), etc.
- Uncertainty Principle: T-F resolution tradeoff
- Non-stationary processes
  - Flares
  - Trends and Modulations
  - Statistical change-points
- Instantaneous Frequency
- Local vs. Global structure
- Interference (cross-terms in bi-linear representation)

Time-Frequency/Time-Scale Analysis (Temps-Fréquence) Patrick Flandrin http://perso.ens-lyon.fr/patrick.flandrin/publis.html; A Wavelet tour of Signal Processing (Une Exploration des Signaux en Ondelettes) Stéphane Mallat

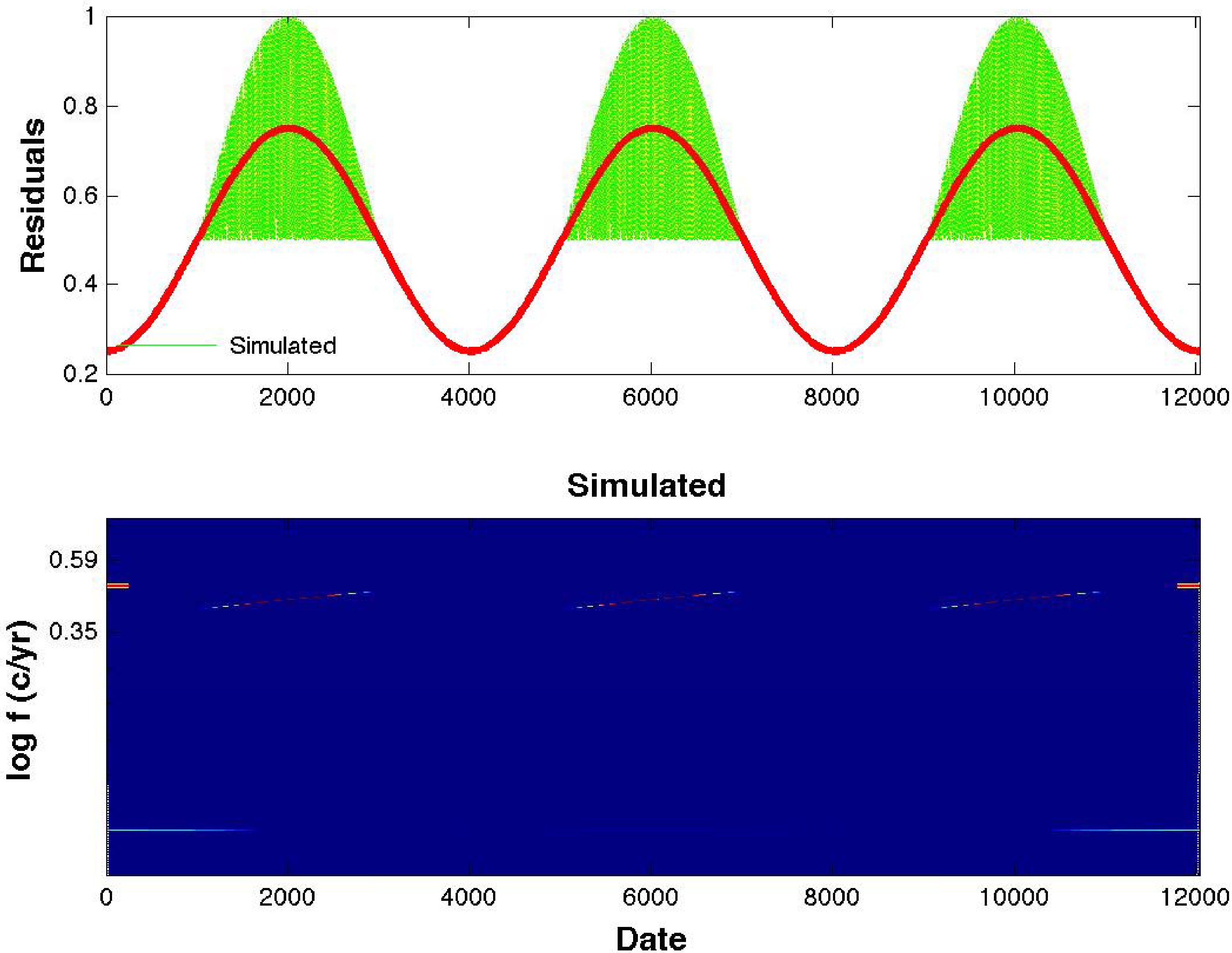
# SYNCHROSQUEEZING

***A consistent and invertible time-frequency analysis tool that can identify and extract oscillating components (of time-varying frequency and amplitude)***

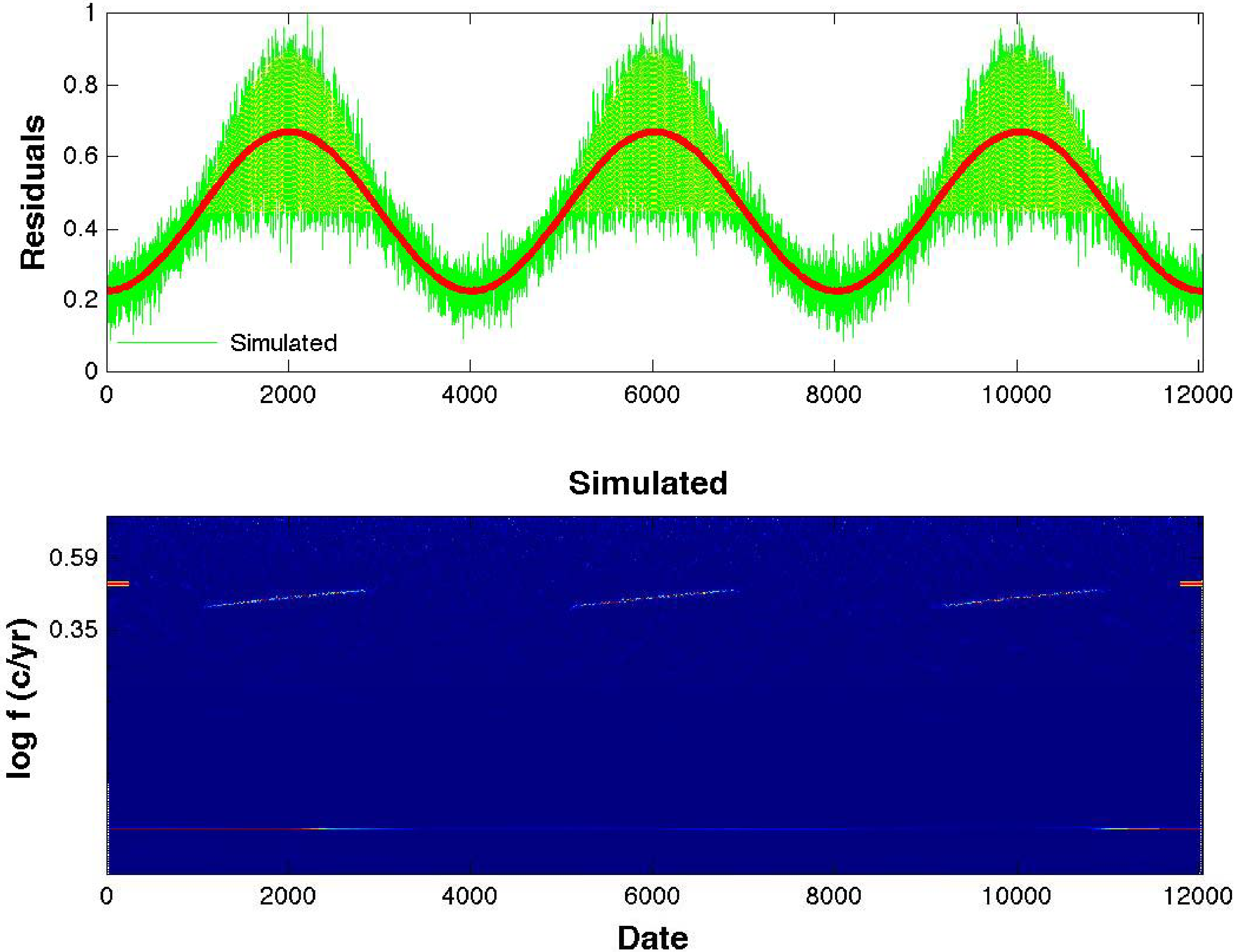
E. Brevdo, N.S. Fučkar, G. Thakur, and H-T. Wu, *The Synchrosqueezing algorithm: a robust analysis tool for signals with time-varying spectrum*, Submitted, 2011 (arXiv id: 1105.0010)

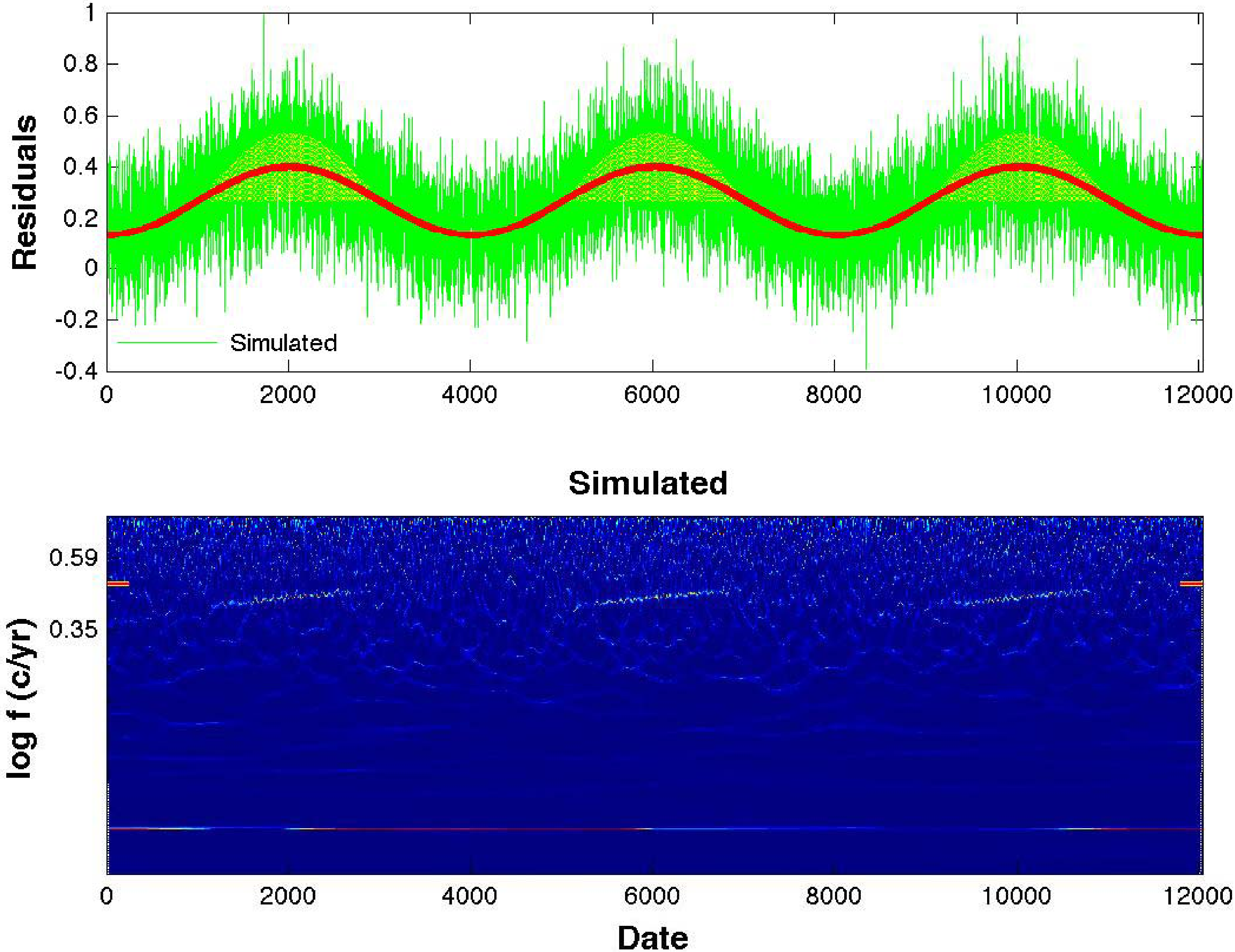
I. Daubechies, J. Lu, and H.-T. Wu, *Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool*, Applied and Computational Harmonic Analysis, 2010.

Matlab tools: http://www.math.princeton.edu/~ebrevdo/synsq/

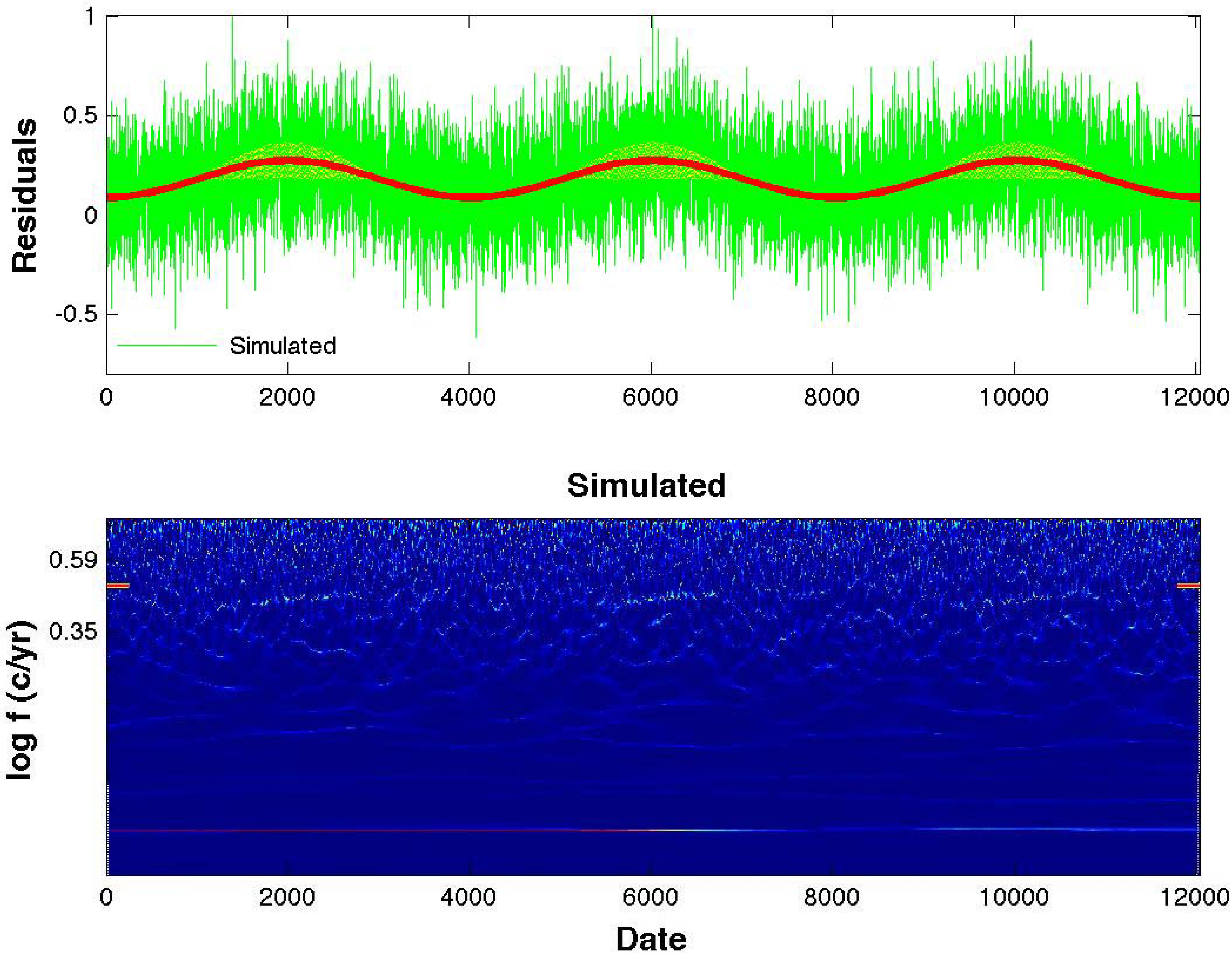


Residuals
1
0.8
0.6
0.4
0.2
0
2000
4000
6000
8000
10000
12000
Simulated
Simulated
log f (c/yr)
0.59
0.35
0
2000
4000
6000
8000
10000
12000
Date

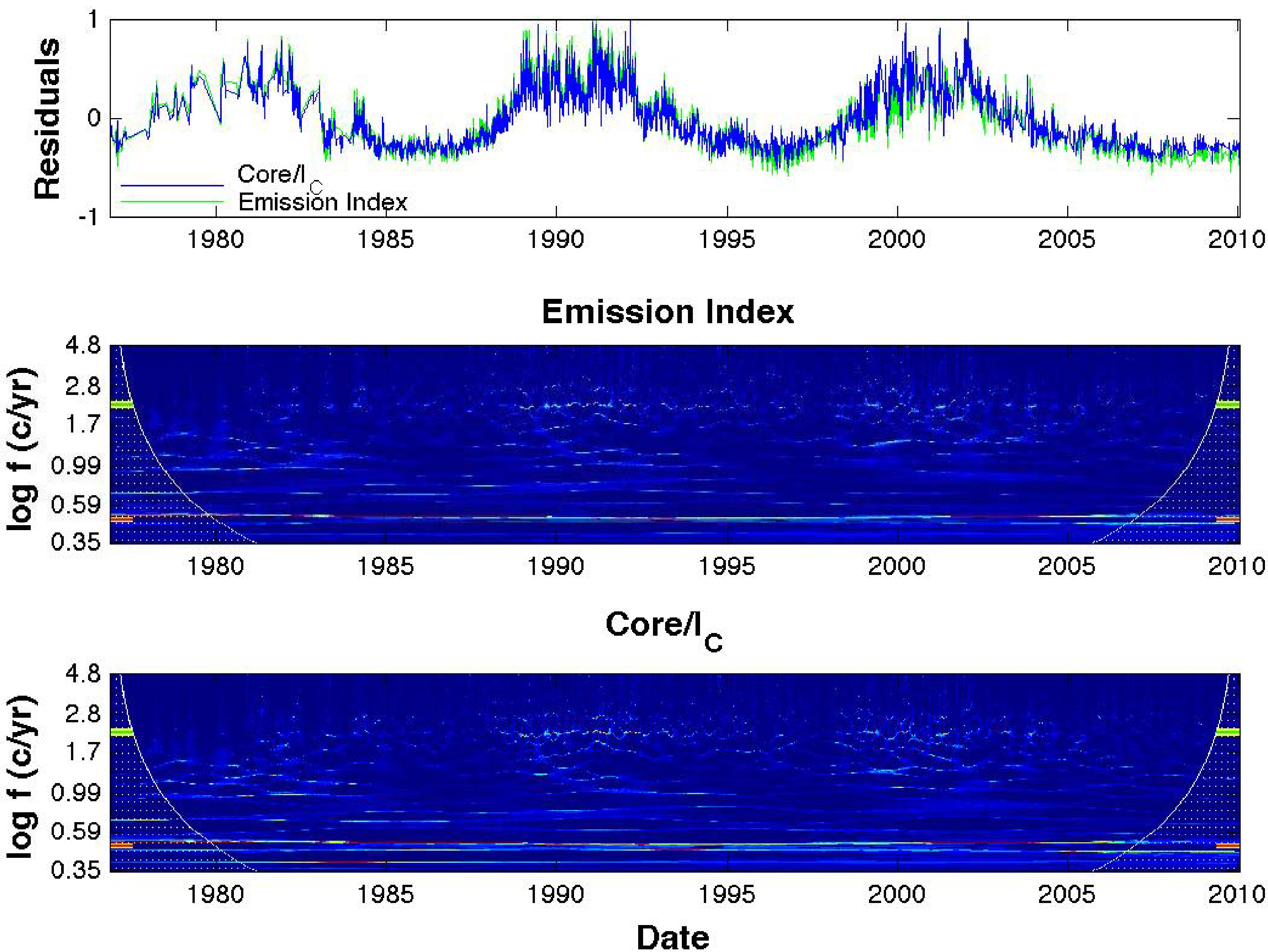

Residuals
Simulated
0
0.2
0.4
0.6
0.8
1
0
2000
4000
6000
8000
10000
12000
Simulated
log f (c/yr)
0.59
0.35
Date

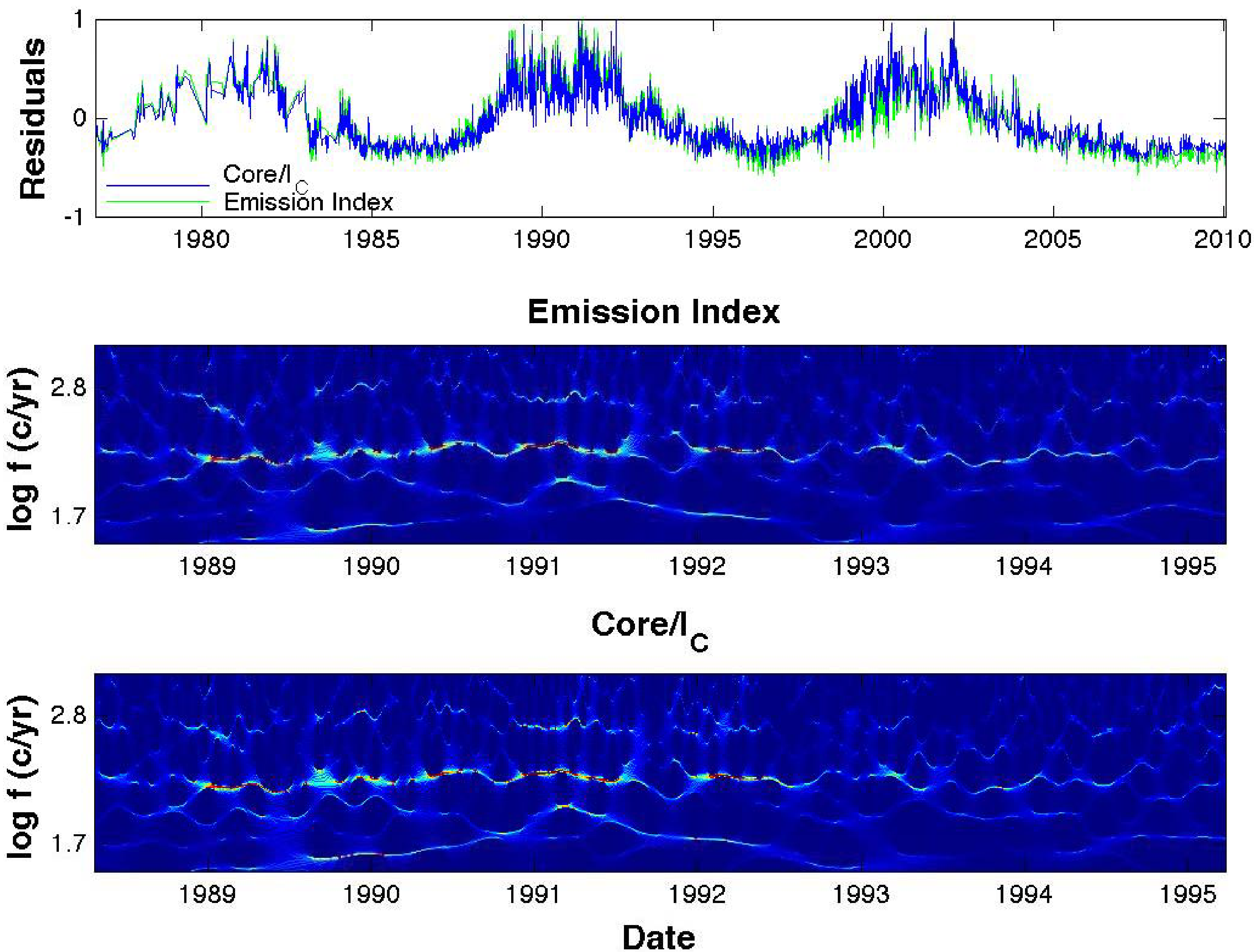


Residuals
Simulated
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
2000
4000
6000
8000
10000
12000
Simulated
log f (c/yr)
0.59
0.35
Date



Residuals
1
0.5
0
-0.5
Simulated
0
2000
4000
6000
8000
10000
12000
Simulated
log f (c/yr)
0.59
0.35
0
2000
4000
6000
8000
10000
12000
Date


Residuals
Core/I_C
Emission Index
1
0
-1
1980
1985
1990
1995
2000
2005
2010
Emission Index
log f (c/yr)
4.8
2.8
1.7
0.99
0.59
0.35
Core/I_C
Date

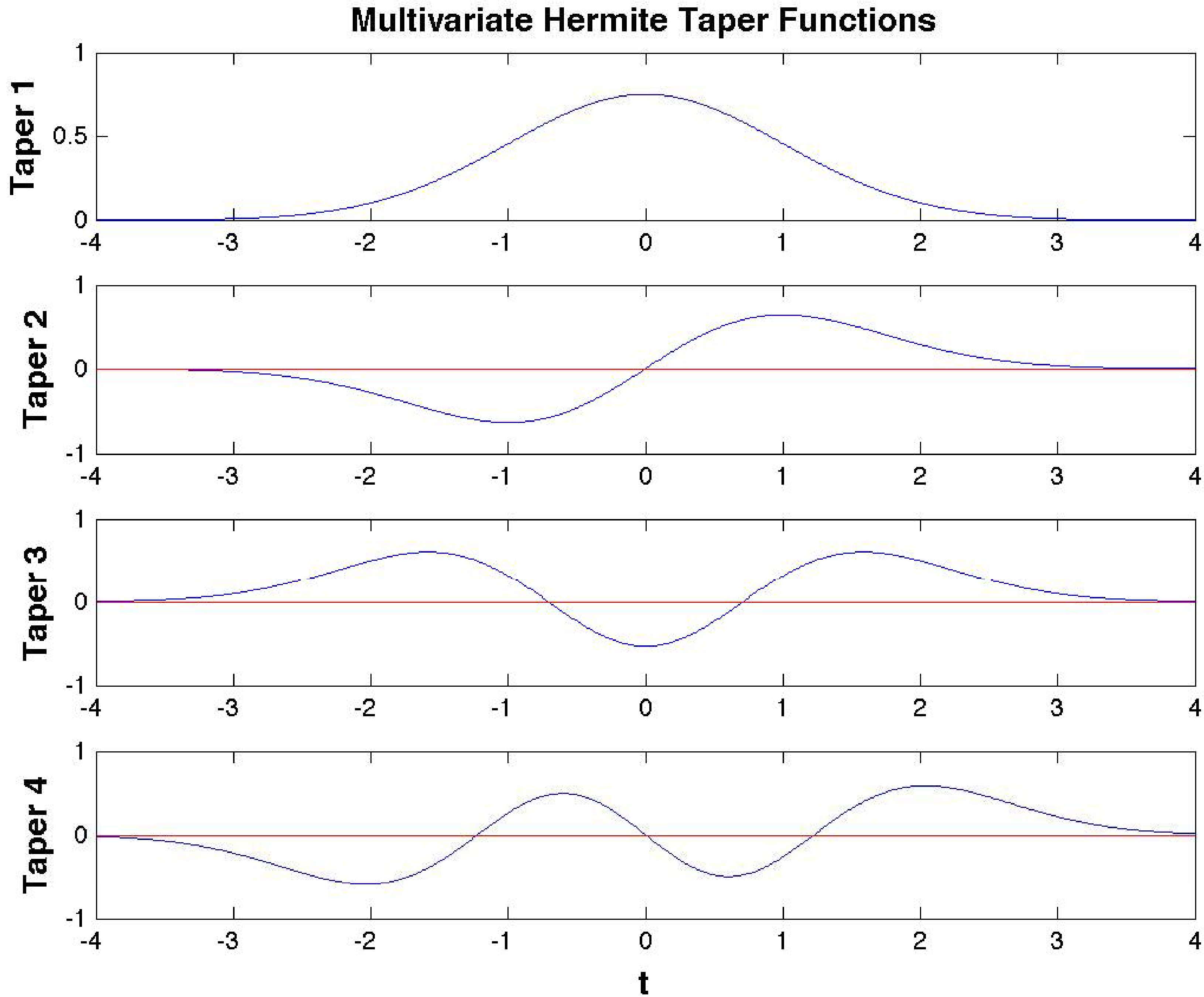


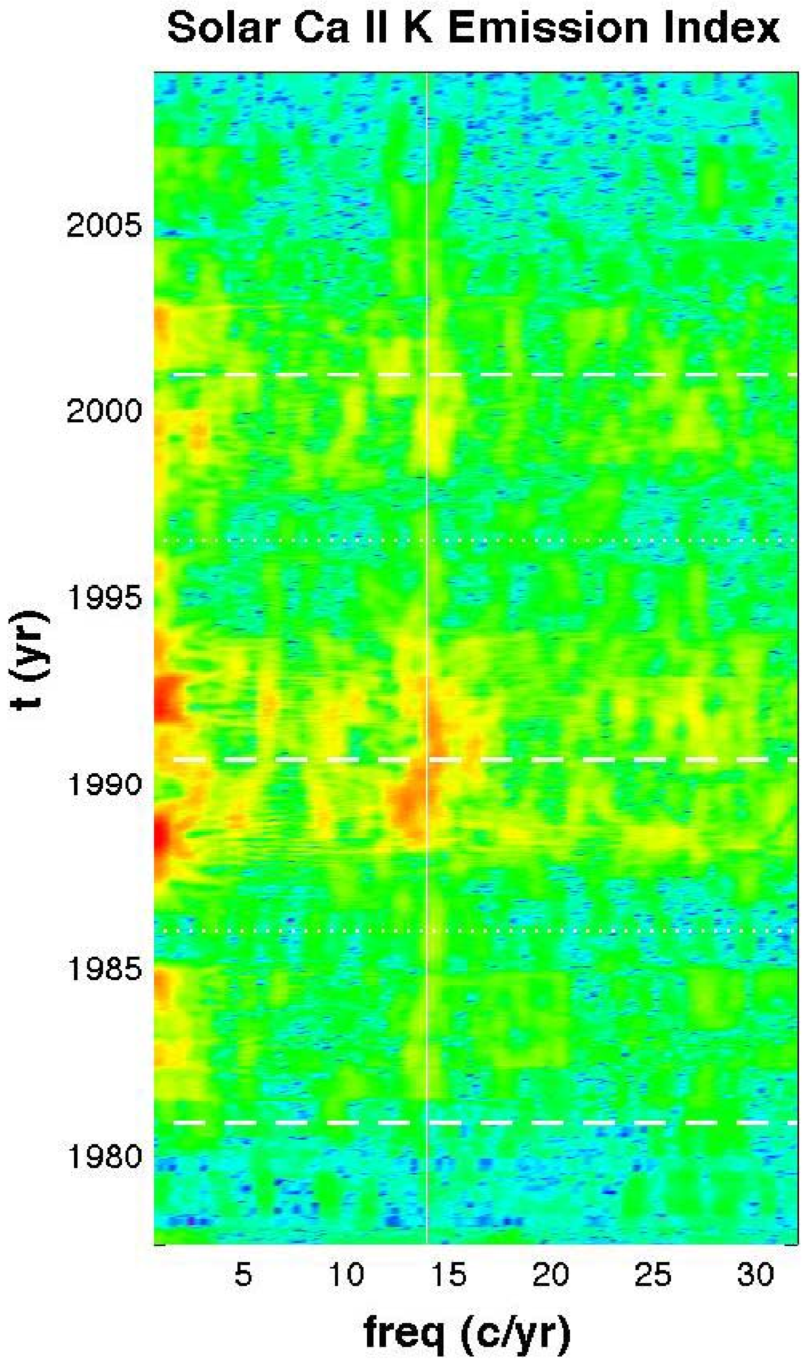
Residuals
Core/I_C
Emission Index
1
0
-1
1980
1985
1990
1995
2000
2005
2010
Emission Index
log f (c/yr)
2.8
1.7
1989
1990
1991
1992
1993
1994
1995
Core/I_C
log f (c/yr)
2.8
1.7
1989
1990
1991
1992
1993
1994
1995
Date

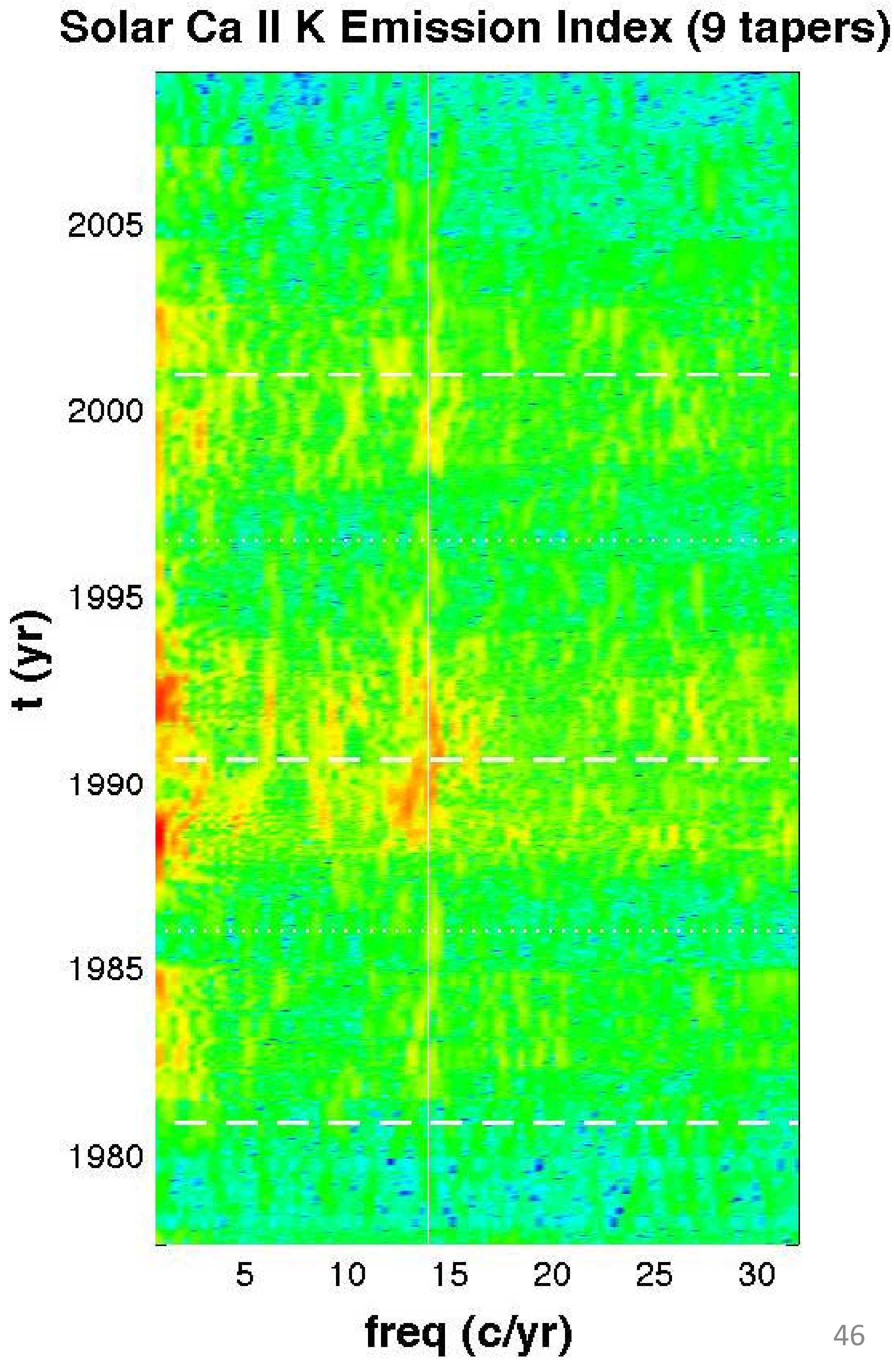
# Multi-taper Analysis (Thomson 1982)

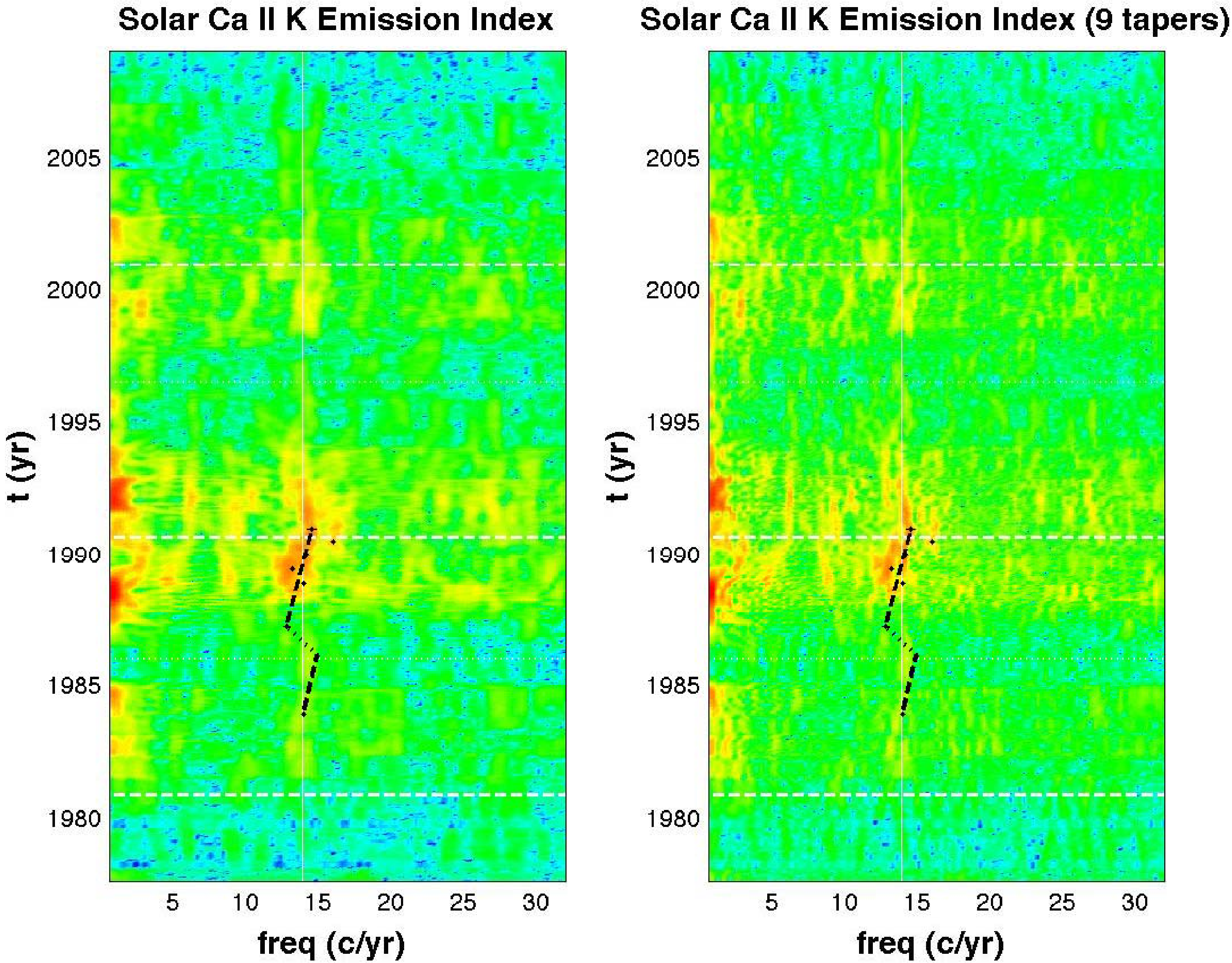
- Tapers (windows) reduce sidelobe leakage = bias
- Incomplete use of data ➔ loss of information
- Multitapers recover this information
- Leakage minimization = eigenvalue problem
  - Eigenfunctions: efficient window functions
  - Eigenvalues
    - measure effectiveness
    - determine how many terms to include

*Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*, Don Percival and Andrew Walden (1993)

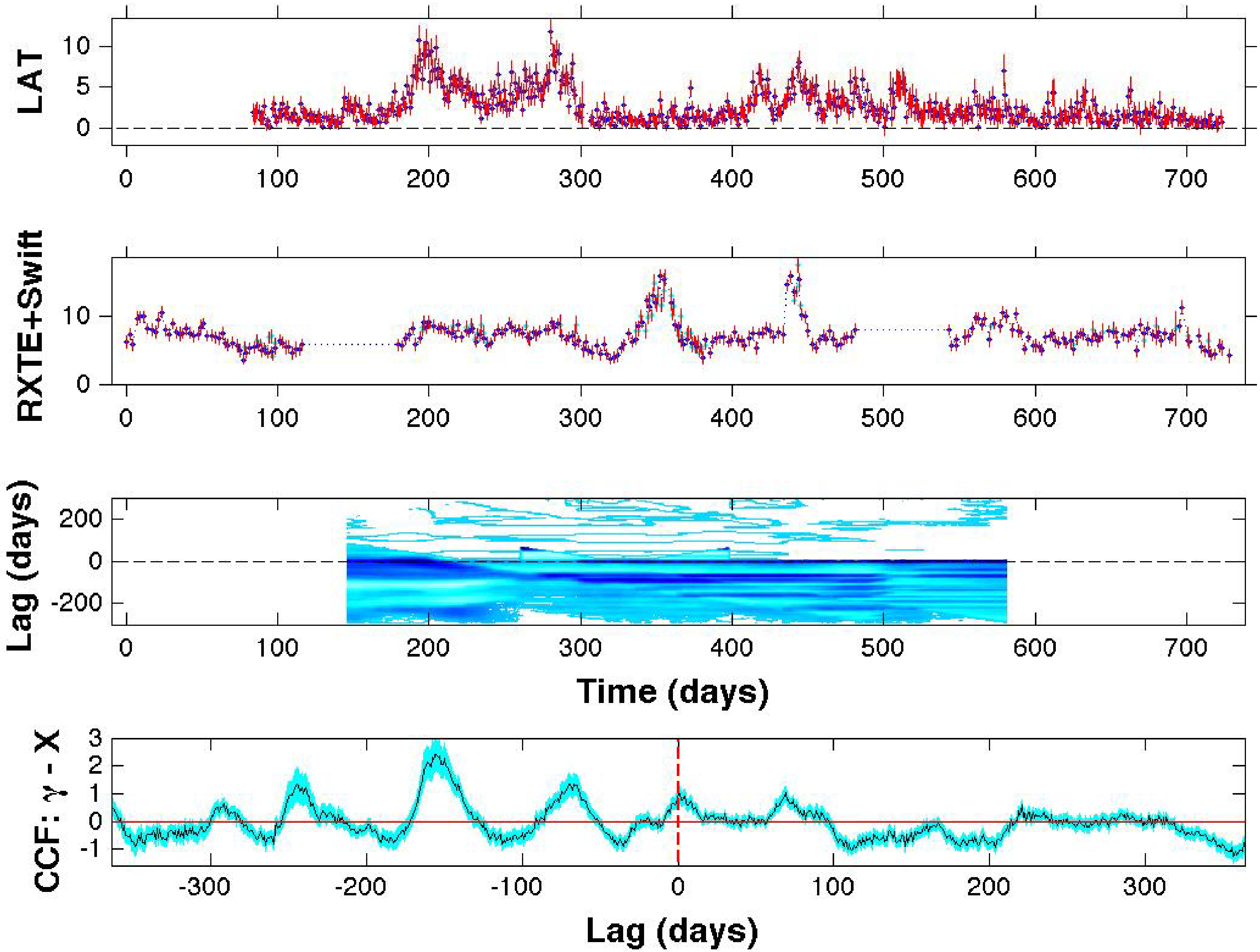

Multivariate Hermite Taper Functions
Taper 1
Taper 2
Taper 3
Taper 4
1
0.5
0
-1
-4
-3
-2
-1
0
1
2
3
4
t


Solar Ca II K Emission Index
2005
2000
1995
1990
1985
1980
t (yr)
5
10
15
20
25
30
freq (c/yr)


Solar Ca II K Emission Index (9 tapers)
2005
2000
1995
1990
1985
1980
t (yr)
5
10
15
20
25
30
freq (c/yr)



Solar Ca II K Emission Index
Solar Ca II K Emission Index (9 tapers)
t (yr)
2005
2000
1995
1990
1985
1980
5
10
15
20
25
30
freq (c/yr)



LAT
RXTE+Swift
Lag (days)
Time (days)
CCF: γ - X
Lag (days)

| Function | Domain | Range | Auto- | Cross- | Physical Interp |
|---|---|---|---|---|---|
| Bayesian blk. Light Curve | Time | Flux | ✓ | ✓ multivar. BB | Flares, events etc. |
| Scatter Plot | Flux 1 | Flux 2 | | ✓ | Dependency (not just cor.) |
| Correlation | Lag | $\langle X^2 \rangle$ $\langle XY \rangle$ | ✓ | ✓ | Correlated behavior/lags |
| Spectrum | Frequency | Power | ✓ | ✓ | Periodicity 1/f noise … |
| | | Phase | ✓ | ✓ | Shifts, lags |
| Structure | Lag | $\langle X^2 \rangle$ $\langle XY \rangle$ | ✓ | ✓ | Correlated behavior/lags |
| Scalogram | Scale/Time | Power | ✓ | ✓ | Dynamic behavior |
| Scalegram | Scale | Power | ✓ | ✓ | 1/f noise QPOs |
| Distribution | Time/scale/ frequency | Power | ✓ | ✓ | Dynamic behavior |

# Open Problems

- The Arrow of Time
- Bayesian Blocks $O(N^2) \rightarrow O(N)$ or $O(N \log N)$
- Optimal Partitioning on the Circle
- Negative Power Spectrum Estimates
- Phase Spectrum from Edelson and Krolik-style Correlations
- Exploration of Time-Scale/Time-Frequency Variants
- Understand Wold Representation if $C = C(t)$
- Automatic Classification of Time Series

***Handbook of Statistical Analysis of Event Data***

MatLab Code; Documentation; Examples; Tutorial
… funded by the NASA AISR Program

***Advances in Machine Learning and Data Mining for Astronomy***

Editors: Michael Way, Jeff Scargle
Kamal Ali, Ashok Srivastava

Chapman and Hall, an imprint of CRC Press
(a division of Taylor and Francis)
Randi Cohen, Computer Science Acquisitions Editor Data Mining and Knowledge Discovery series

# *Back-ups*

# Practical Suggestions

***(somewhat exaggerated)***

- Study distribution of sample intervals $dt_n = t_{n+1} - t_n$
- Never subtract mean of time series
- Edelson and Krolik CF is the source of all other analysis
- Use self terms in E&K CF to assess observational errors
- Don't confuse: source randomness/observational noise
- $H_0$: AGNs are identical stochastic dynamical systems
- Stationarity is a local property
- Any stationary random process is exactly shot noise (random pulses; the Wold Decomposition Theorem)
- Linearity is a physical property, not one of time series
- Do not bin data