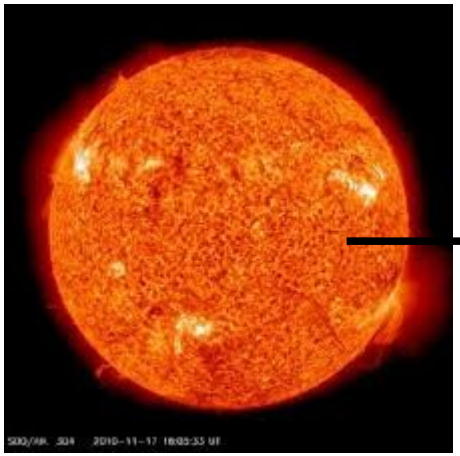
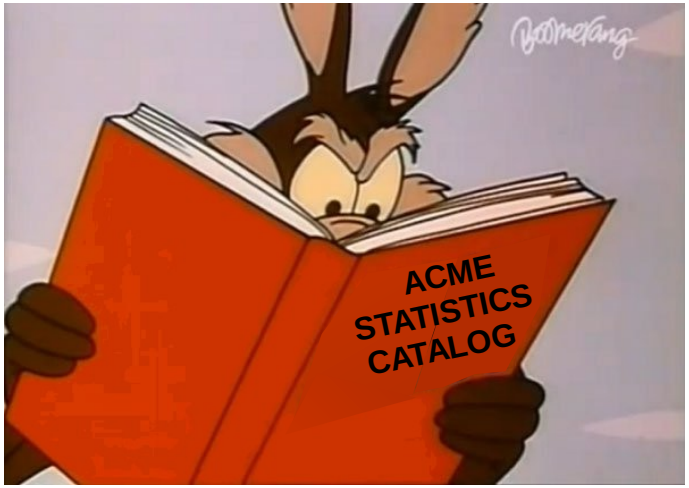
Statistical Issues for the Solar Neutrino Researcher



ν



Scott Oser
University of British Columbia
PhyStat-ν 2016
September 20, 2016

Outline

I will avoid spending much time on issues covered by Michael Smy at PhyStat-Nu in Kashiwa.

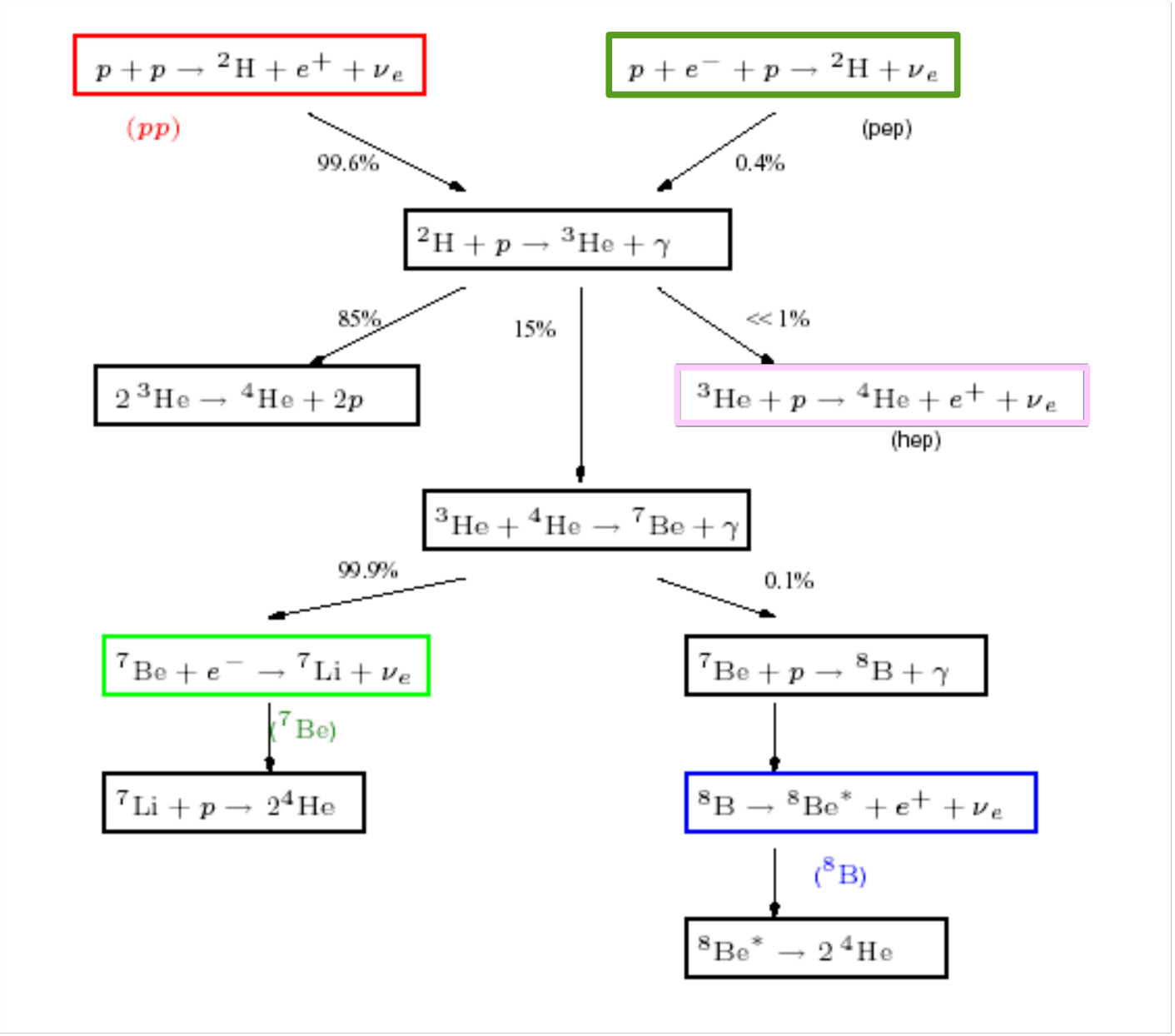
Instead I'll focus on a few areas:

- Issues in maximum likelihood signal extraction
- Statistical issues in time variability analyses

My background is from SNO, and apologies in advance for choosing most of my examples from that experiment, but it's what I know best. I'll try to indicate the general applicability of the methods.

Michael Smy, “Solar Neutrinos”, PhyStat-ν in Kashiwa, May 2016,
http://indico.ipmu.jp/indico/getFile.py/access?contribId=35&sessionId=2&resId=0&materialId=slides&confId=82

The *pp* fusion chain



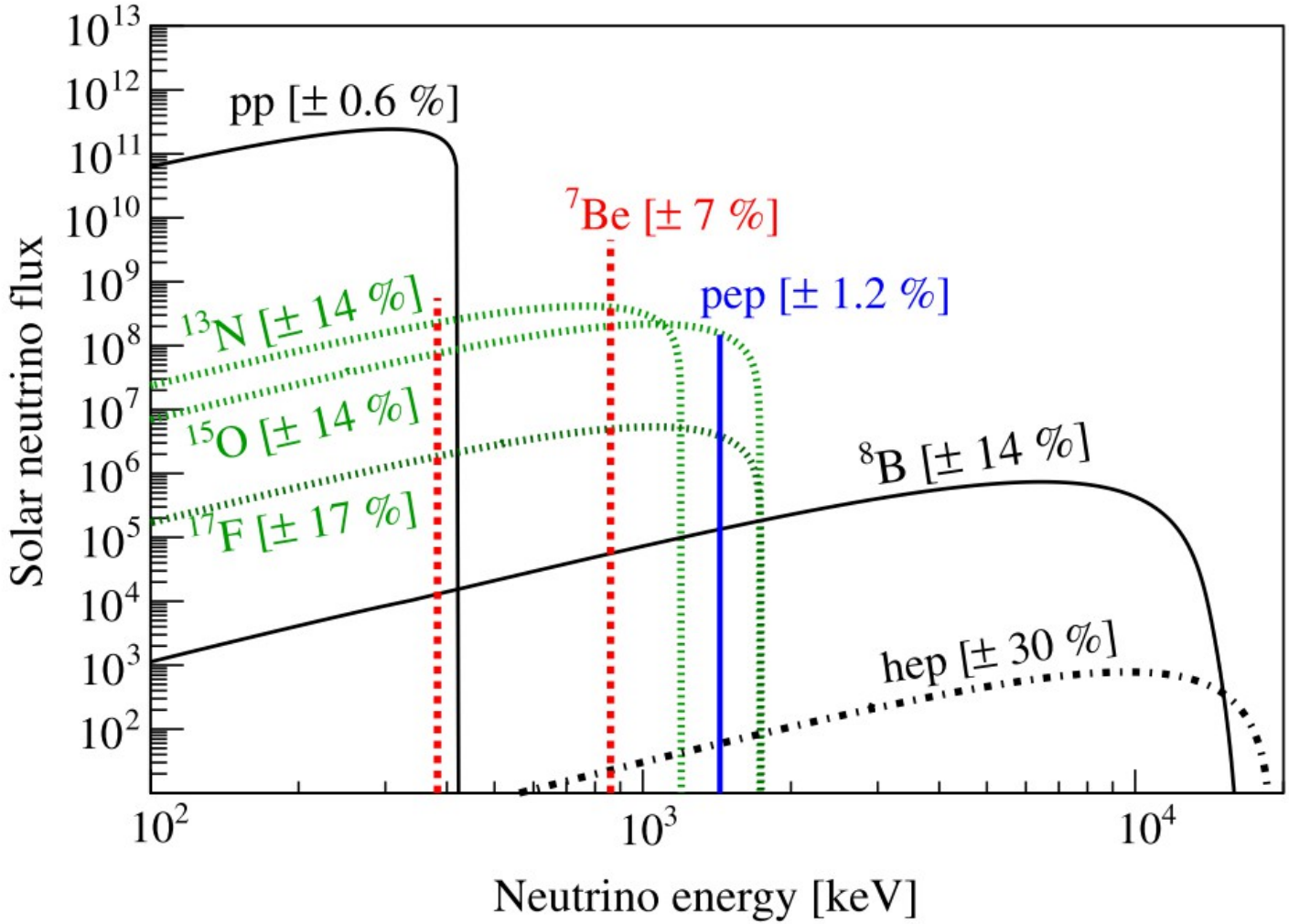
Dominant source of energy production in the Sun

Solar neutrino energy spectra

Shape of spectra determined by straightforward nuclear physics.

Rates determined by astrophysics.

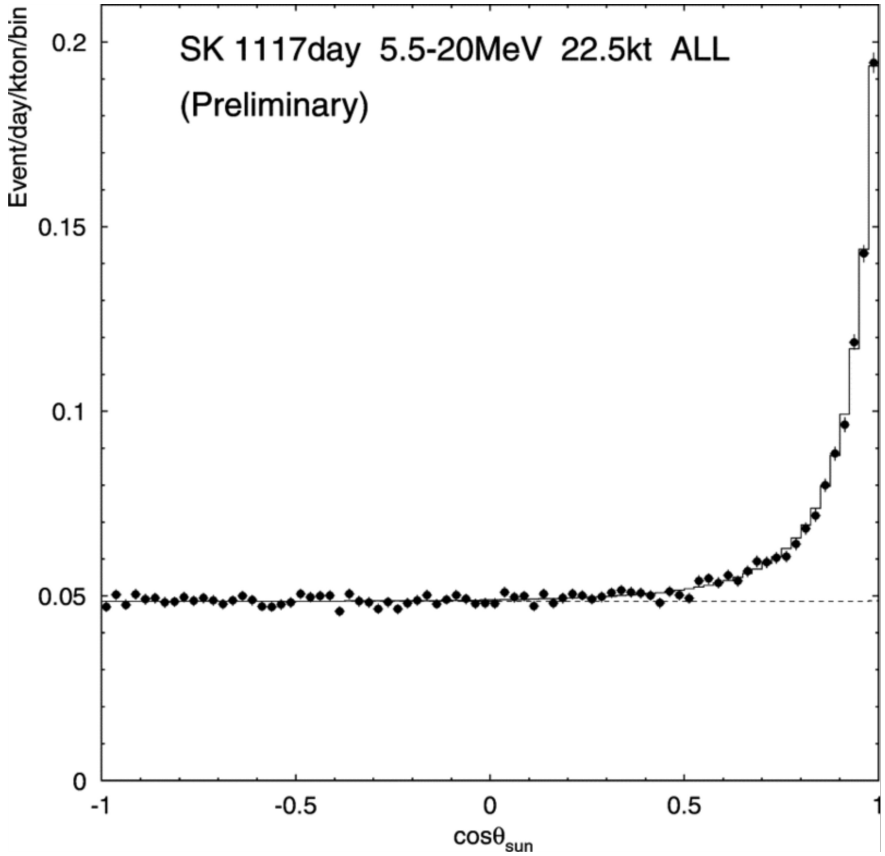
Oscillations modify the observed spectra and rates.



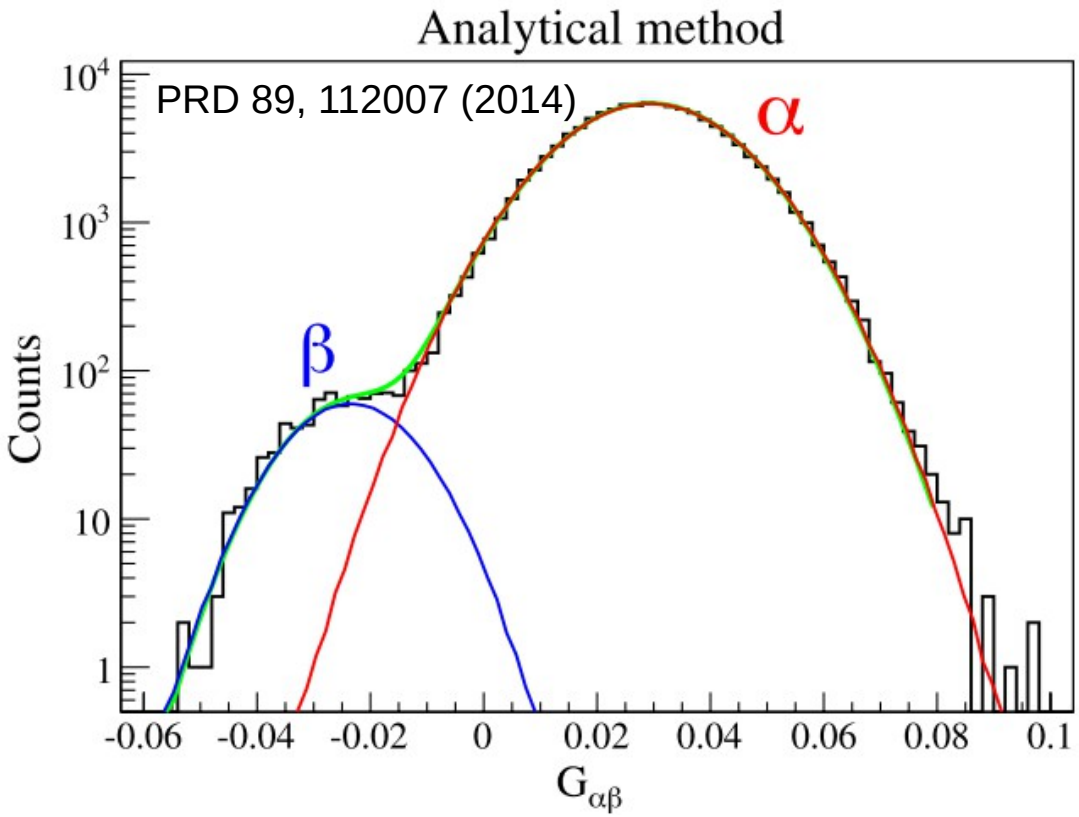
Detection techniques

Technique	Reaction	Threshold	Flavour Sensitivity	Energy measured	Real time detection
Radio-chemical	CC on ^{37}Cl	0.81 MeV	ν_e	N	N
	CC on ^{71}Ga	0.23 MeV	ν_e	N	N
Water Cherenkov	ES on H_2O	~3.5 MeV	Mostly ν_e	Y	Y
	CC on D_2O	~3.5 MeV	ν_e	Y	Y
	NC on D_2O	2.2 MeV	All	N	Y
Scintillation	ES on C_3H_4	~0.06 MeV	Mostly ν_e	Y	Y

Telling signal from background



Angular correlation of Super-K events with solar direction to separate ES peak from radioactive background



"Gatti parameters", based on scintillation light time profile, used to separate α decays from electron events in Borexino

Most experiments use extended maximum likelihood fits to extract neutrino rates from large radioactive backgrounds.

Telling signal from signal (SNO)

Radioactive backgrounds were less of a problem at SNO, due to its cleanliness and energy threshold. More critical was distinguishing between three signal reactions, with different flavour sensitivity.

This generally cannot be done on an event-by-event basis. But fits to distributions can disentangle them.

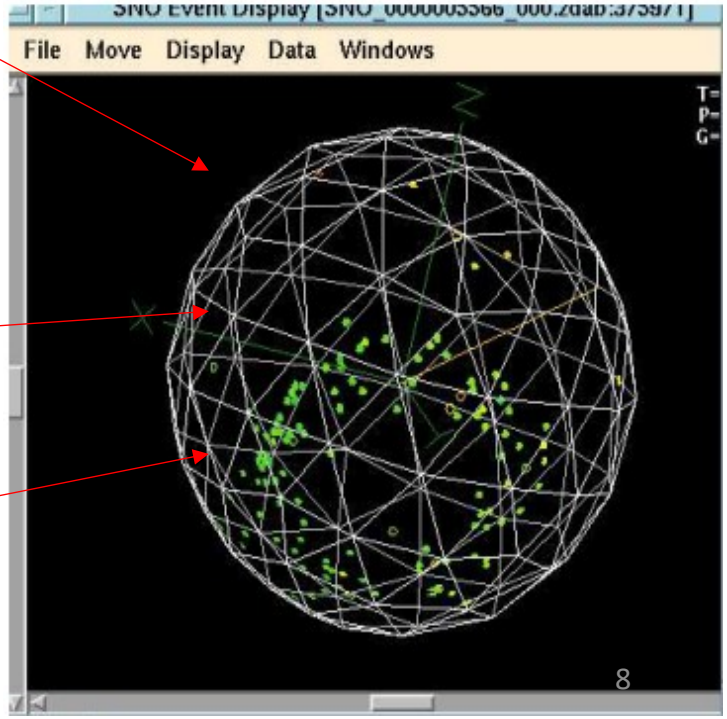
Solar Neutrino Interactions in Pure D_2O

NC

$$\nu_x + d \rightarrow p + n$$

$$n + d \rightarrow t + \gamma \ (6.25\ \text{MeV})$$

In each case, the end result is a a single Cherenkov ring in SNO



CC

$$\nu_e + d \rightarrow p + p + e$$

ES

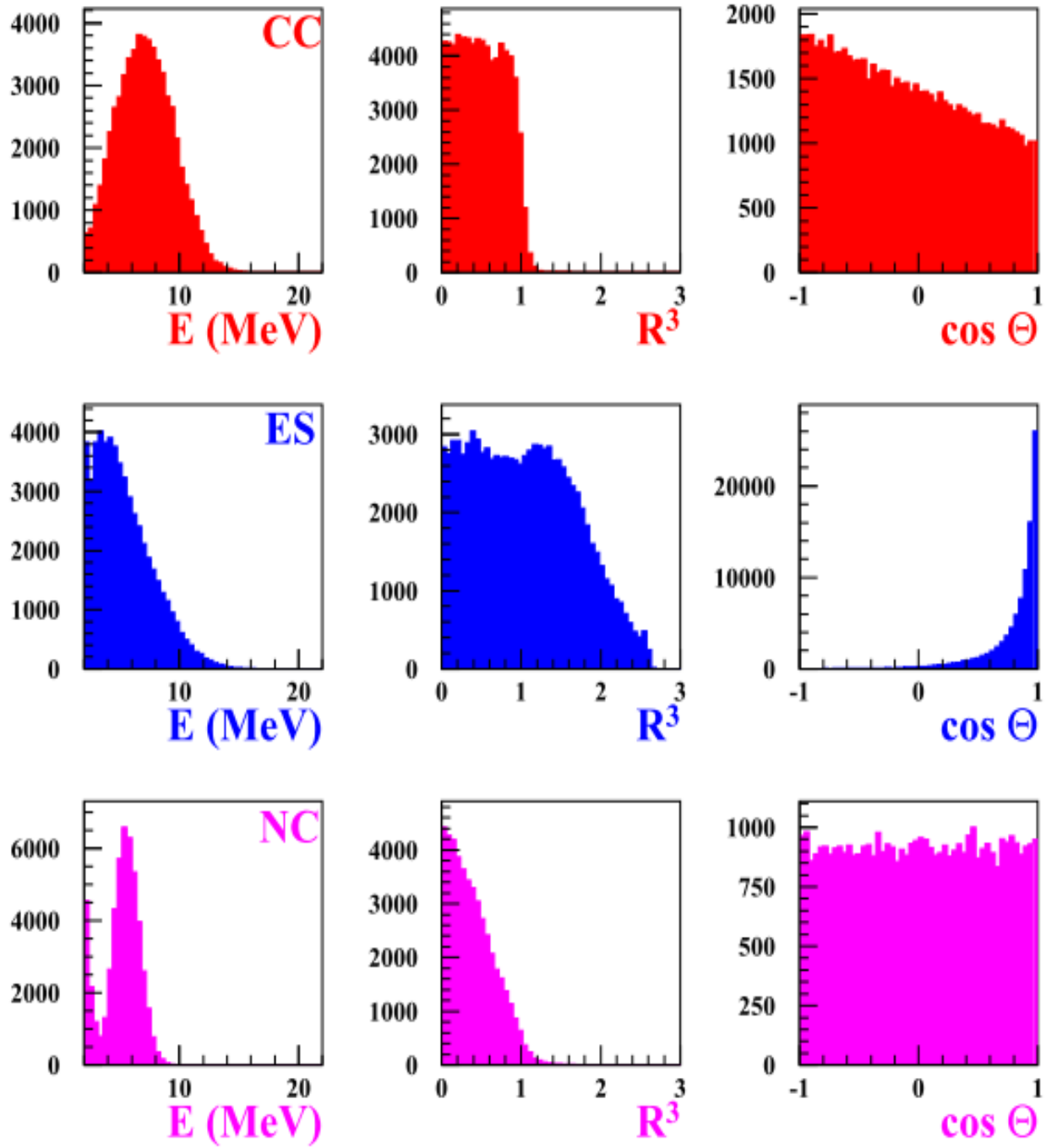
$$\nu_x + e \rightarrow \nu_x + e$$

Distinguishing CC, ES, and NC reactions

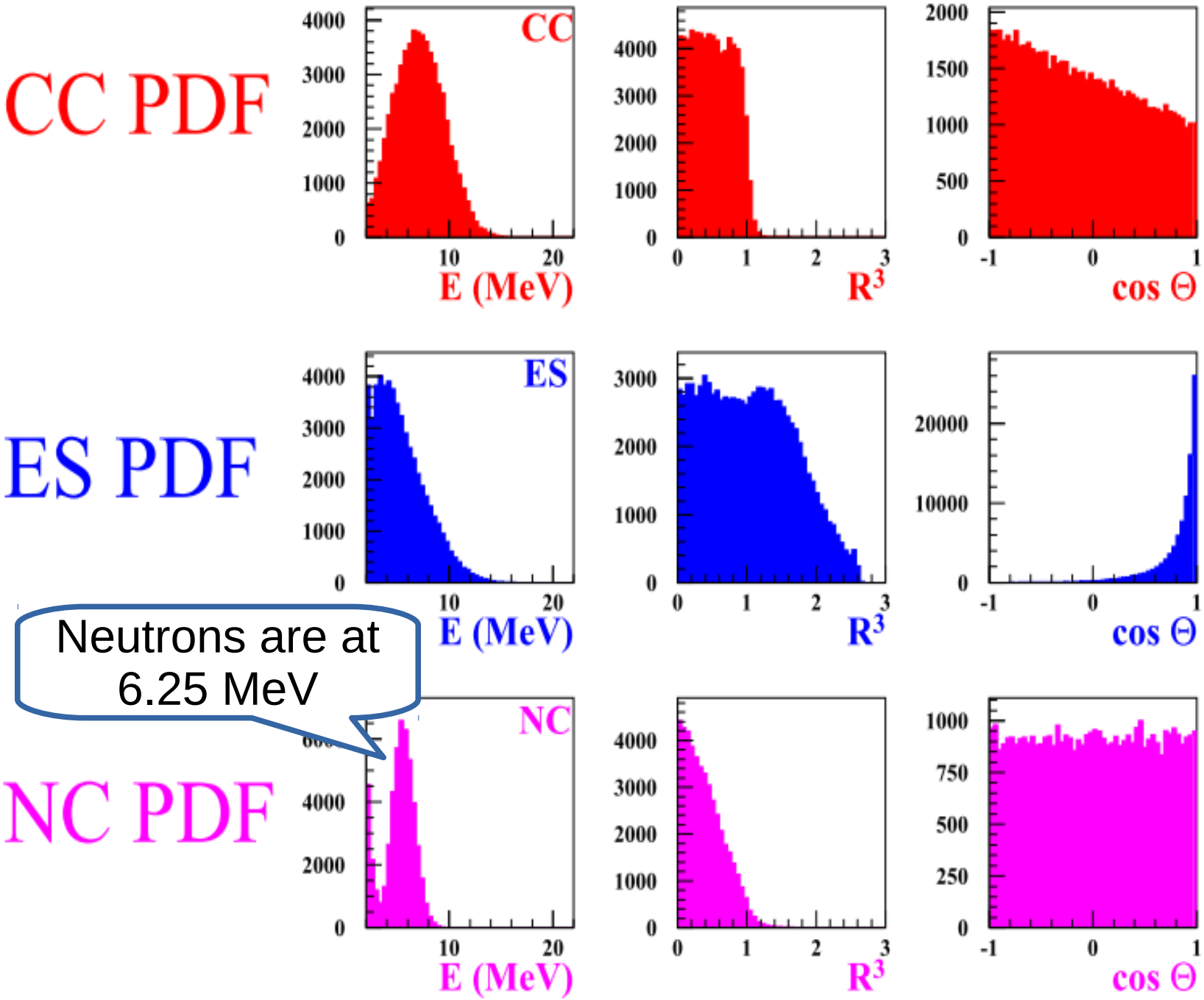
CC PDF

ES PDF

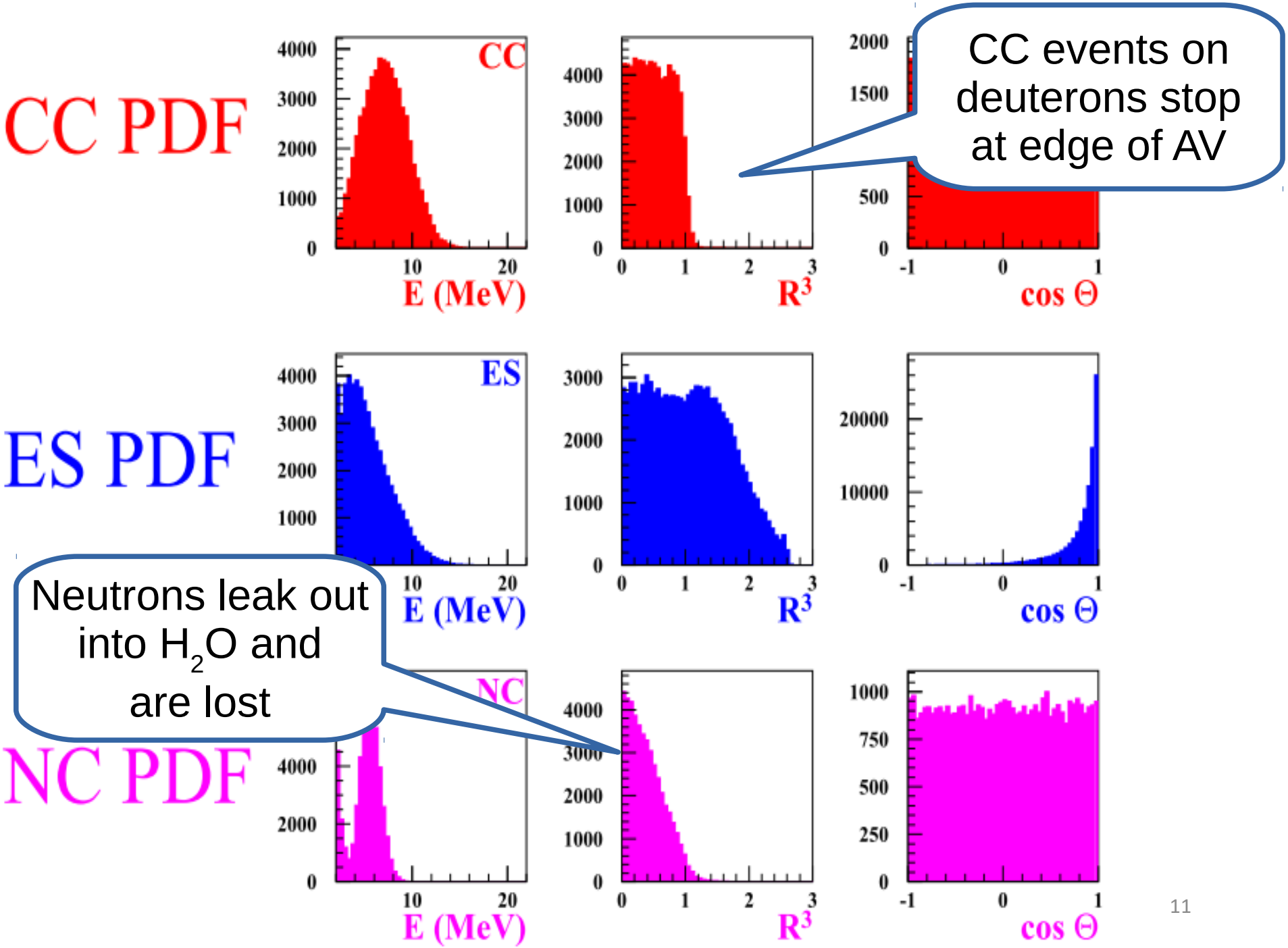
NC PDF



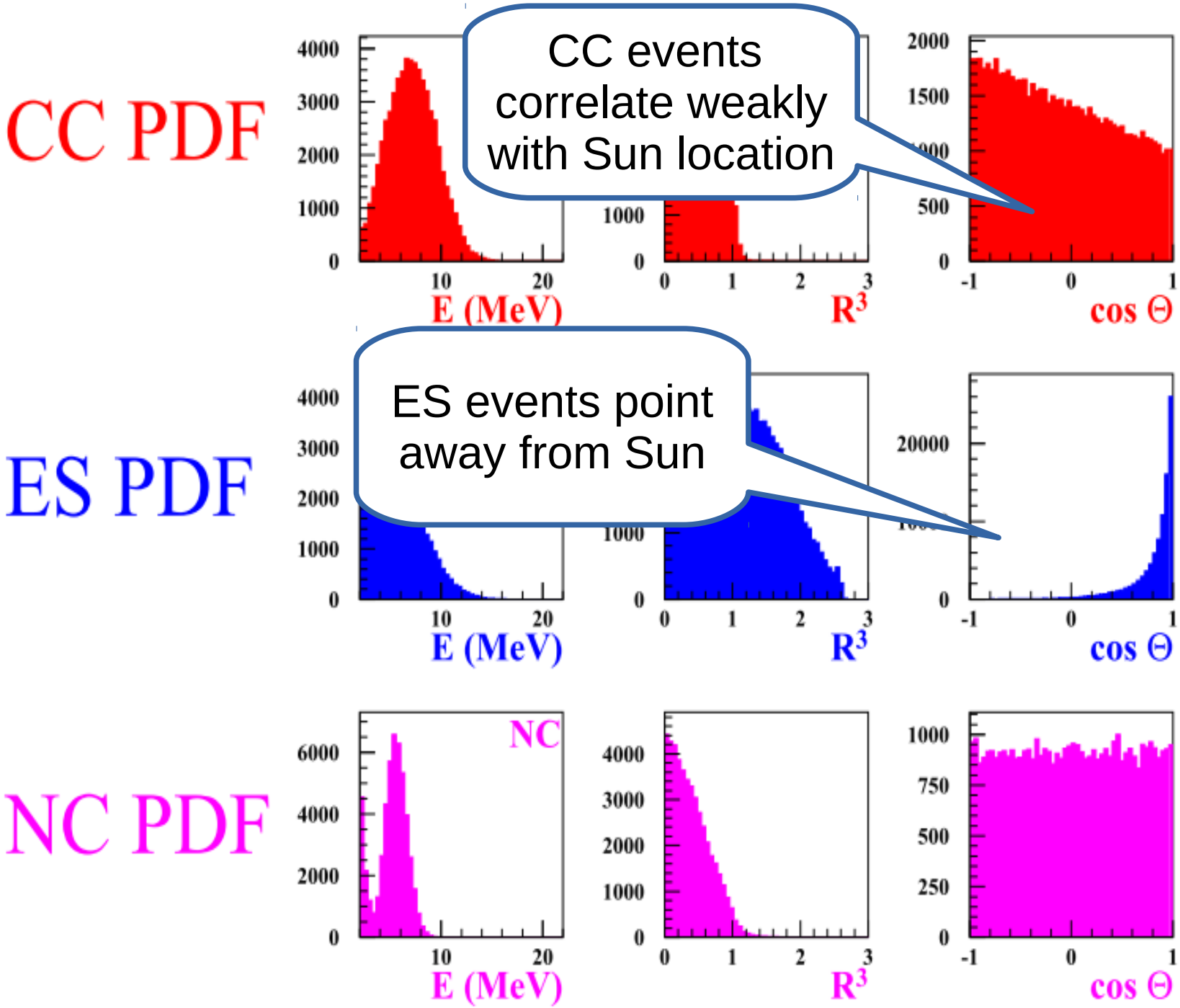
Distinguishing CC, ES, and NC reactions



Distinguishing CC, ES, and NC reactions



Distinguishing CC, ES, and NC reactions



SNO's Simplest Extended ML Fit

$$-\ln L = \sum_{i=CC,ES,NC} N_i - \sum_{j=1}^{\text{events}} \ln\left[\sum_{i=CC,ES,NC} N_i f_i\left(E_j, R_j^3, \theta_j\right)\right]$$

Fit for N_{CC}, N_{NC}, N_{ES} using three 3D PDFs in (E, R^3, θ). Or reparametrize in terms of $\Phi(^8B)$ and constant $P(\nu_e \rightarrow \nu_e)$.

PDFs are assembled from Monte Carlo events, and in practice are formed by binning Monte Carlo events in 3D. It's important to capture correlations between the three observables, especially E and θ for elastic scattering events.

Key limitation: does not allow spectral distortions, but sufficient for energy-independent oscillation probability.

Allowing for spectral distortions

$$\sum_{k}^{energy\ bins}\left\{\sum_{i=CC,ES,NC} N_{i,k} - \sum_{j=1}^{\text{events}_k} \ln\left[\sum_{i=CC,ES,NC} N_{i,k} f_{i,k}\left(R_j^3,\theta_j\right)\right]\right\}$$

Add separate likelihood terms for each energy bin, and fit for number of CC and ES events in each bin. Constrain $N_{NC,k}$ to follow neutron energy PDF shape.

For active neutrino oscillations the ES spectrum is not independent from the CC spectrum. We can instead reparametrize $N_{CC,k}$ and $N_{ES,k}$ in terms of $\Phi(^8B)$ and an energy-dependent $P(\nu_e \rightarrow \nu_e)$.

SNO extracted both binned CC/ES spectra, and also parametrized fits for $P(\nu_e \rightarrow \nu_e) = A + BE_\nu + CE_\nu^2$

Floating Systematics

Signal PDFs depend on detector response parameters such as energy scale, energy resolution, and vertex resolution. These can be accounted for by making the PDFs be functions of these parameters, and including the parameters, along with a prior constraint term, in the likelihood. For example, if energy scale is uncertain, we can shift Monte Carlo event energies when calculating PDFs:

$$E_{MC} \rightarrow \alpha E_{MC}$$

Then the likelihood is modified to include the dependence of the PDFs on α:

$$N - \sum_i f(E_i) \rightarrow N - \sum_i f(E_i, \alpha) + \frac{1}{2}\left(\frac{\alpha - 1}{\sigma_\alpha}\right)^2$$

SNO's initial publications did not use this, but instead “shifted and refit” for each systematic, then added results in quadrature. This overestimated total error by about 20%!

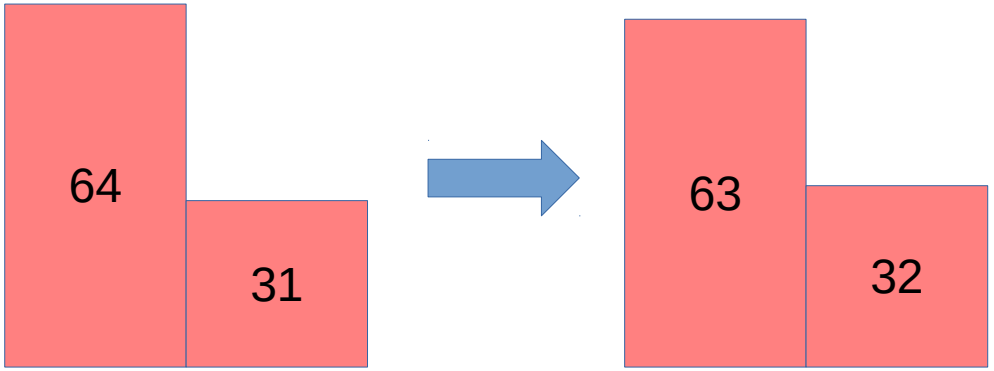
Computational complexities of floating systematics

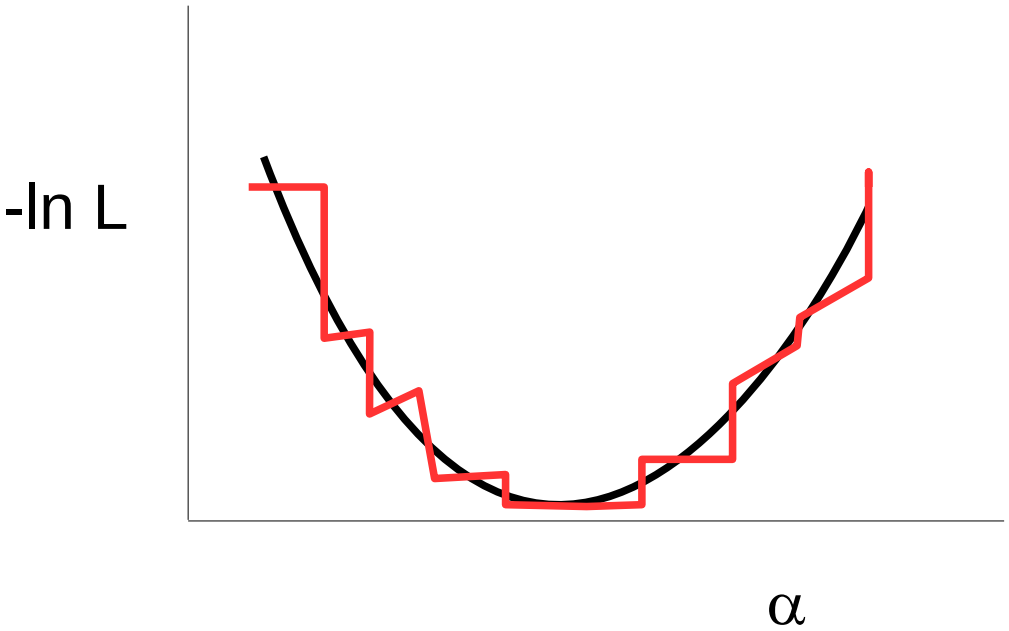
PDFs are no longer constant in fit, but must be recalculated at each step of minimization. This is a big CPU suck. Various approaches:

- Keep all Monte Carlo events in memory, rebuild PDFs at each step of fit. Slowest solution, but safest.
- Linearly parametrize dependence of each PDF on each systematic, and use analytic approximations (effectively a first order Taylor series) to calculate modifications to PDF without keeping Monte Carlo events in memory. Risks missing non-linear effects.
- Create a giant covariance matrix between all bins, and include in fit. Forces a Gaussian approximation.

An issue with binned MC PDFs and floating systematics

If PDFs are built from binning MC events, then a small shift in a detector response parameter can discontinuously change the number of events in a bin, as events are shifted across bin boundaries.





Likelihood surface becomes discontinuous as a function of some fit parameters. Most fitters, especially those based on calculating gradients, go bonkers. The likelihood surface may have many local minima as well.

Markov Chain Monte Carlo

Treat likelihood function as a probability distribution to be sampled from, using the Metropolis algorithm: Propose new point y from x_n.

$$x_{n+1} = y \text{ with probability } p = \min\left(1, \frac{L(y)}{L(x)}\right)$$
$$= x_n \text{ with probability } 1-p$$

Advantages: Avoids needing to minimize any functions. Insensitive to discontinuities in likelihood function. Can be easily parallelized by running independent chain on multiple CPUs. Does not require Gaussian assumptions on any variables.

Gives posterior distribution---Bayesian in flavour, if not in practice.

Disadvantages: Computationally taxing. Not convenient for classical confidence intervals (Neyman construction) [although some will view this as an advantage!]

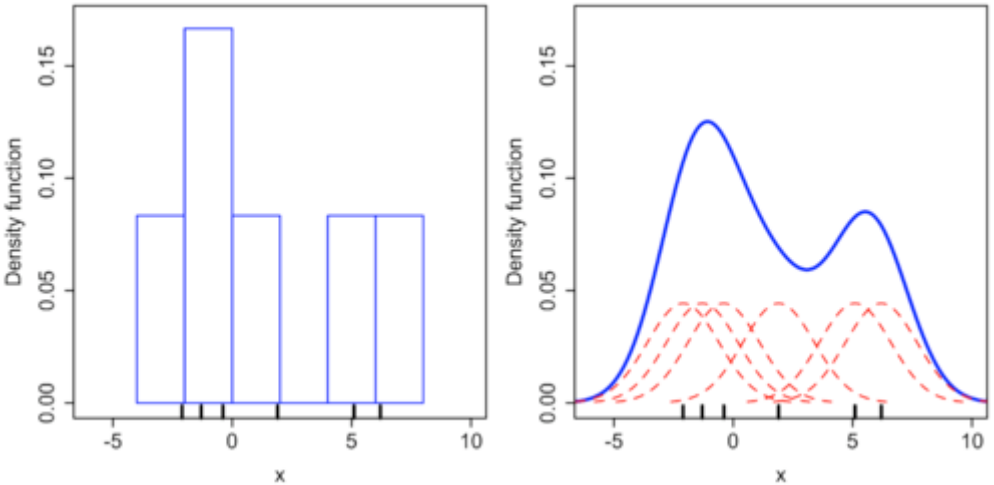
Kernel PDFs

Build an analytic approximation to PDF from MC events:

$$\hat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\sqrt{2\pi}} e^{(x-t_i)^2/2h^2}$$

Here h, called the bandwidth, can be chosen to be different for each event, and can depend on the local density of events.

Advantages: analytic, smooth PDFs can easily be differentiated, and work well with standard fitters. PDFs vary smoothly under transformations of Monte Carlo events.

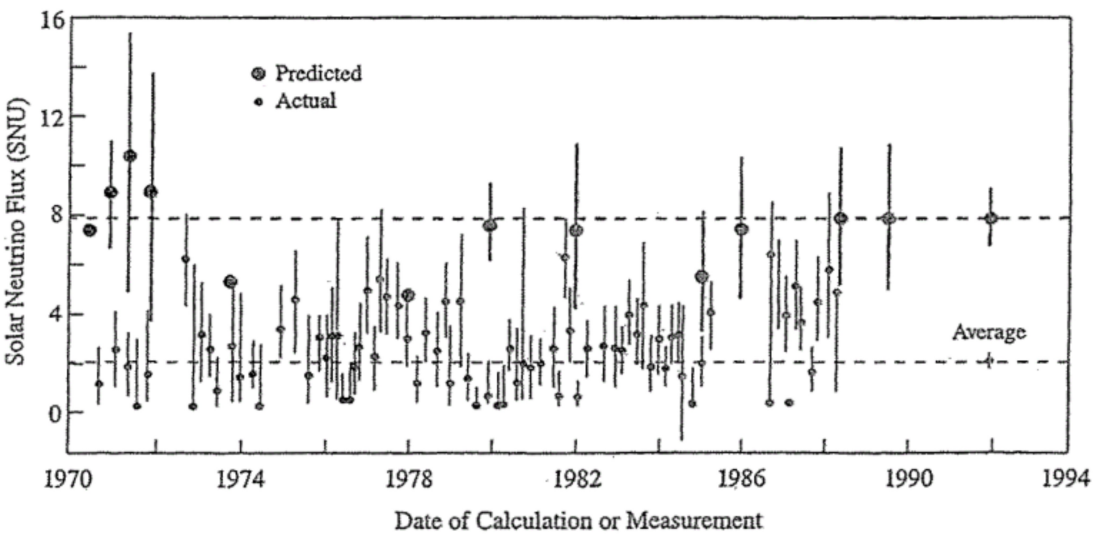


Disadvantages: Results are sensitive to choice of h, which requires tuning. Method is still computationally burdensome.

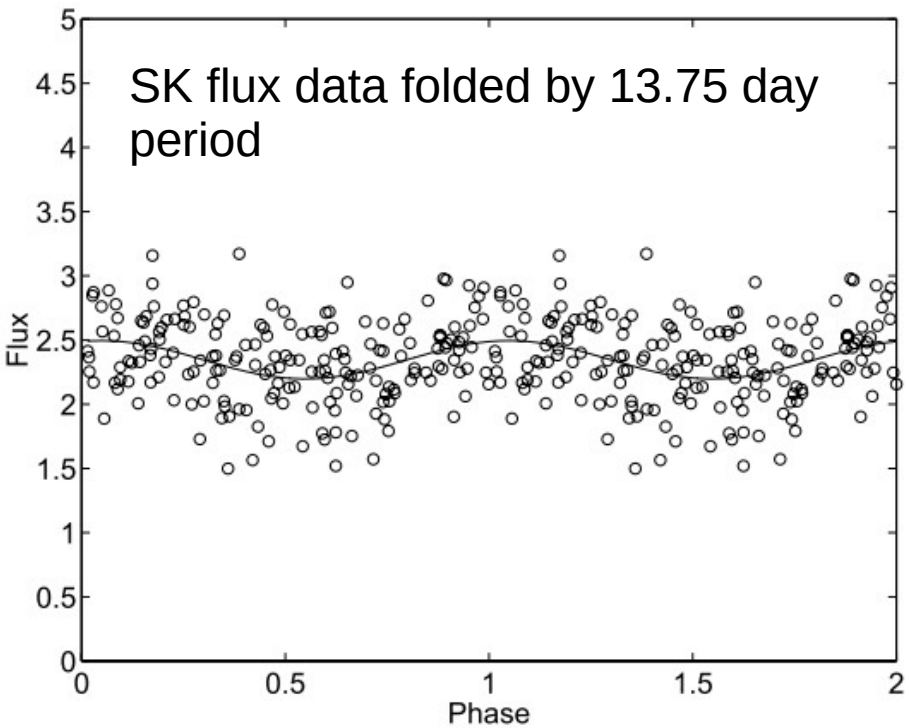
Time Variability Intro

Various groups have re-analyzed solar neutrino rates as a function of time, claiming to find correlations with various solar parameters, such as 11-year solar cycle, solar rotation, r-mode oscillation frequencies. New physics has been implicated.

These claims have generally been sharply contested by the experimenters themselves.



Homestake data



P. Sturrock, Ap.J., 594:1102–1107, 2003 Sept 10

Statistical issues for time variability

- Trials factors: searching at thousands of frequencies simultaneously.
- Exposure: to do analysis properly, one must know the exact times when the detectors are live. Experiments have not always made these public.
- Binning: When exactly did events occur in time bins? Especially relevant for searches at faster frequencies.
- Fishing expeditions: very many analysis methods exist---if one doesn't give an answer you like, try another!

ML Method

$$\phi(t) = W(t) N [1 + A \cos(\omega t + \delta)]$$

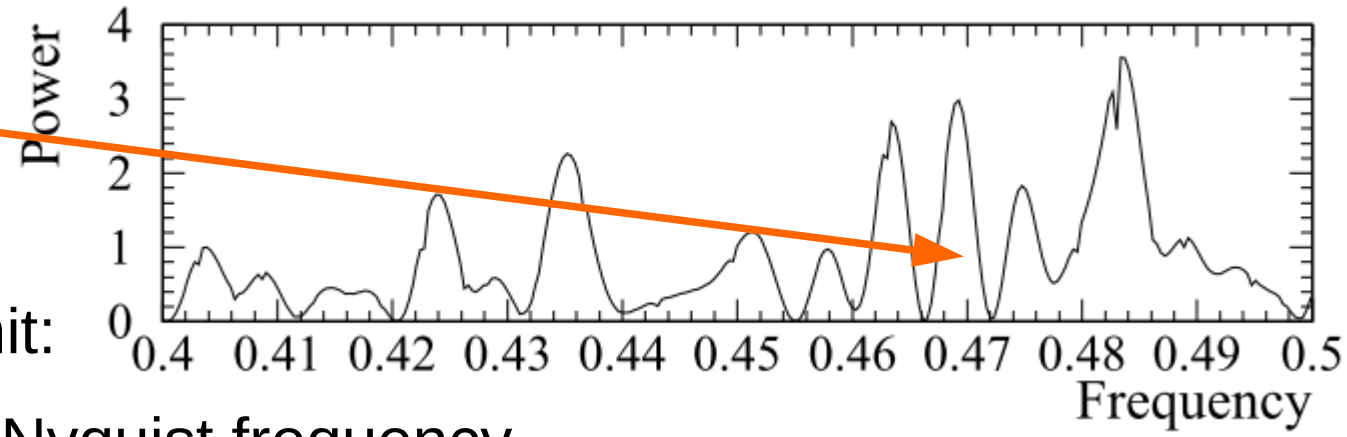
Window function W(t) = 1 if detector live at time t, =0 otherwise. Accounts for exposure history and its time variation.

Do an unbinned likelihood fit at any interesting value of ω, then calculate $S = 2(\ln L_{max} - \ln L_{A=0})$. By Wilks' theorem, under the null hypothesis of A=0 then $P(S > Z) = e^{-Z}$.

However, the value of this test statistic is not independent between different frequencies. Toy Monte Carlo is the best solution to model significance for tests at N different frequencies.

Considerations for frequency scans

There is a frequency resolution given by 1/(duration of data set). Power at frequencies closer together than this will be strongly correlated. Note the finite peak widths!



Maximum frequency limit:

- Evenly binned data: Nyquist frequency
- Unevenly binned data: roll-off in sensitivity at effective Nyquist frequency
- Unbinned analysis: no limit

When scanning over frequencies, one needs to *oversample* by a factor of a few (scan with step sizes smaller than frequency resolution ≈ 1/T). Else significant peaks risk being missed if they are centred in between scanned frequency values.

Lomb-Scargle periodogram: binned data with uneven time intervals

$$P(\omega)=\frac{1}{2\sigma^2}\left(\frac{\left[\sum w_i\left(y(t_i)-\bar{y}\right)\cos\omega\left(t_i-\tau\right)\right]^2}{\sum w_i\cos^2\omega\left(t_i-\tau\right)}+\frac{\left[\sum w_i\left(y(t_i)-\bar{y}\right)\sin\omega\left(t_i-\tau\right)\right]^2}{\sum w_i\sin^2\omega\left(t_i-\tau\right)}\right)$$

$$\tan\left(2\omega\tau\right)=\frac{\sum w_i\sin 2\omega t_i}{\sum w_i\cos 2\omega t_i}$$

Useful when only binned data is available.

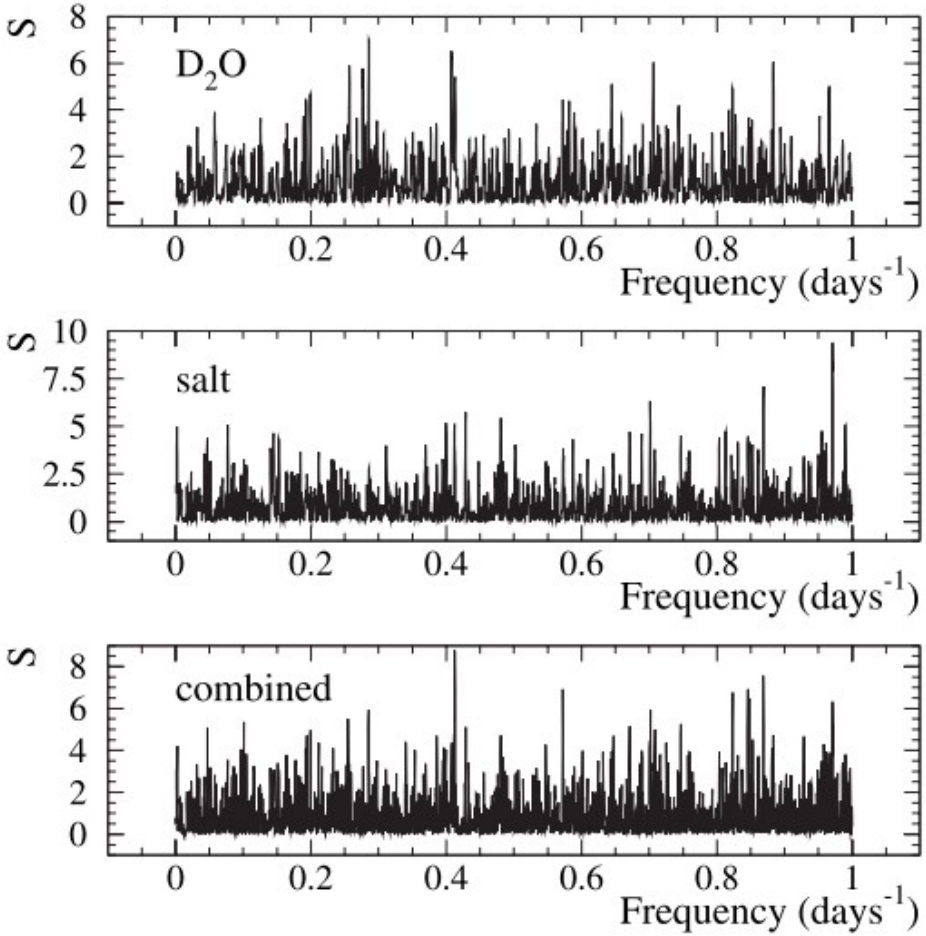
Weights w_i are $1/\sigma_i^2$, & errors assumed to be Gaussian.

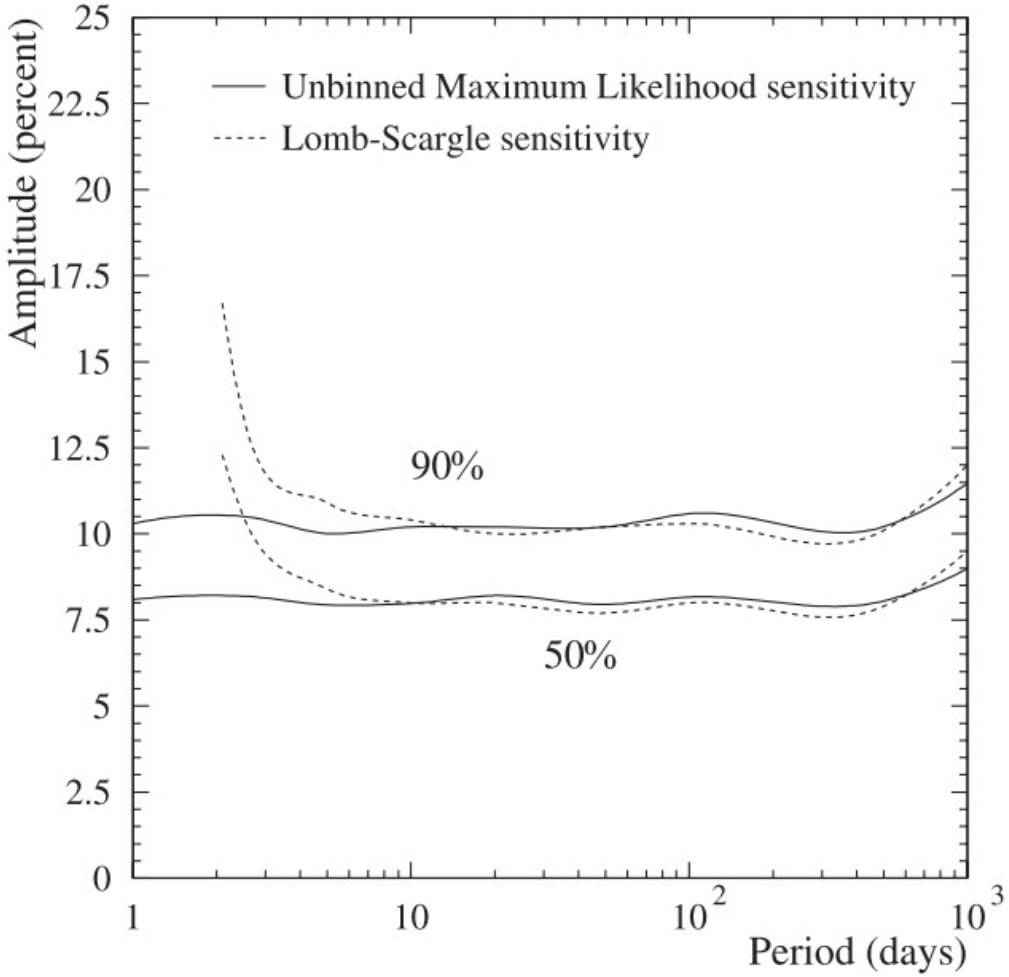
Periodogram is closely related to Fourier transform when data is evenly spaced.

Scargle, J. D. (1982). "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data". Astrophysical Journal. 263: 835

SNO ML periodicity analysis

PHYSICAL REVIEW D **72,** 052010 (2005)





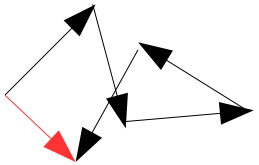
Unbinned analysis has superior sensitivity at shorter periods within a factor of a few of the bin size.

Rayleigh Power

$$z(\omega)=\frac{1}{N}\left[\left(\sum_i \cos\omega t_i\right)^2+\left(\sum_i \sin\omega t_i\right)^2\right]$$

Expected distribution of z sensitive to exposure history. Can calculate it from a random walk model and central limit theorem:

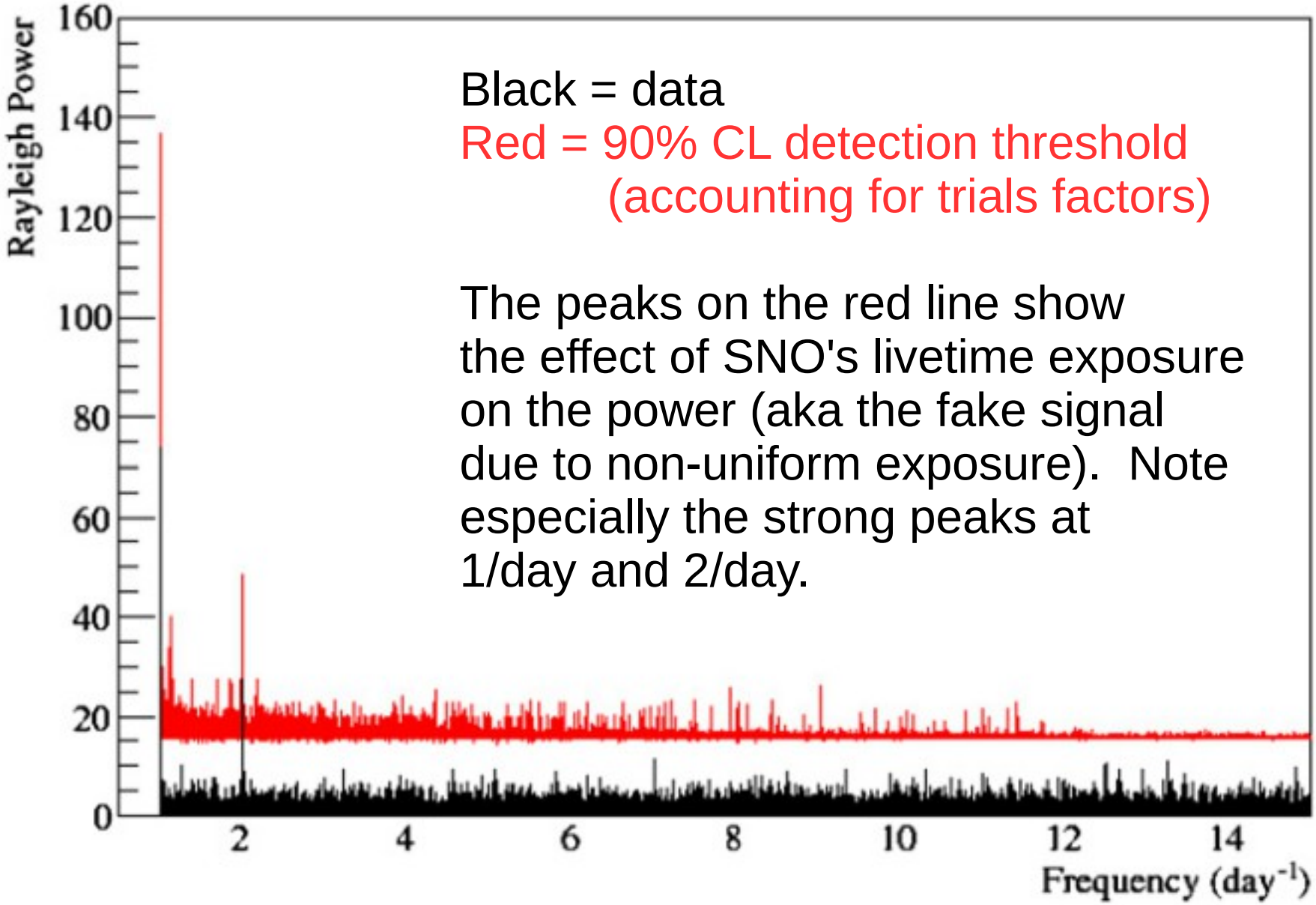
Calculate a phase for each event, and imagine adding the vectors end to end. Periodicity will tend to cause events to align in phase and create a net displacement from the origin: $z = (1/N)[(\Delta X)^2 + (\Delta Y)^2]$.



For a large number of events, ΔX and ΔY are a set of correlated Gaussian variables. Their expected means and covariances can be calculated from their exposures:

$$\mu_x=\frac{1}{T}\sum_{j=1}^{\text{runs}}\int_{t_{start,j}}^{t_{stop,j}} dt\cos\omega t \quad \text{etc.}$$

SNO Rayleigh Power Analysis



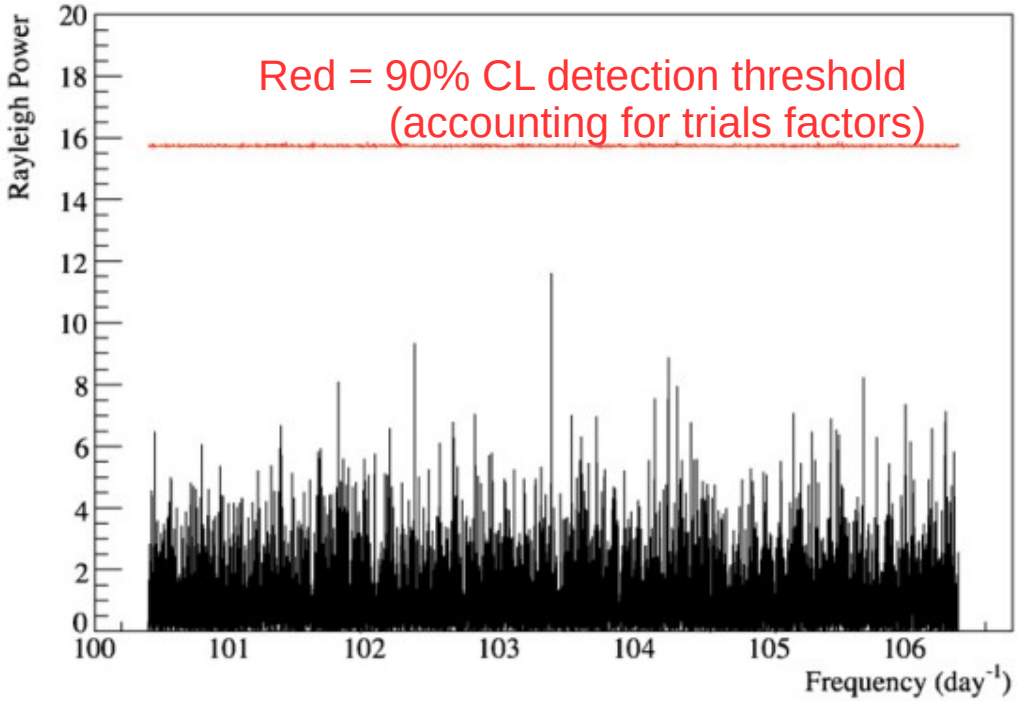
Black = data
Red = 90% CL detection threshold
(accounting for trials factors)

The peaks on the red line show the effect of SNO's livetime exposure on the power (aka the fake signal due to non-uniform exposure). Note especially the strong peaks at 1/day and 2/day.

SNO collaboration, Ap.J. 710:540-548, 2010 Feb 10

High Frequency Search

Rayleigh power method is computationally much faster than unbinned ML method, and suitable for scanning over a large number of frequencies. SNO used this to scan 1.6M frequencies down to 10 minute periods. No sign of any periodicity.



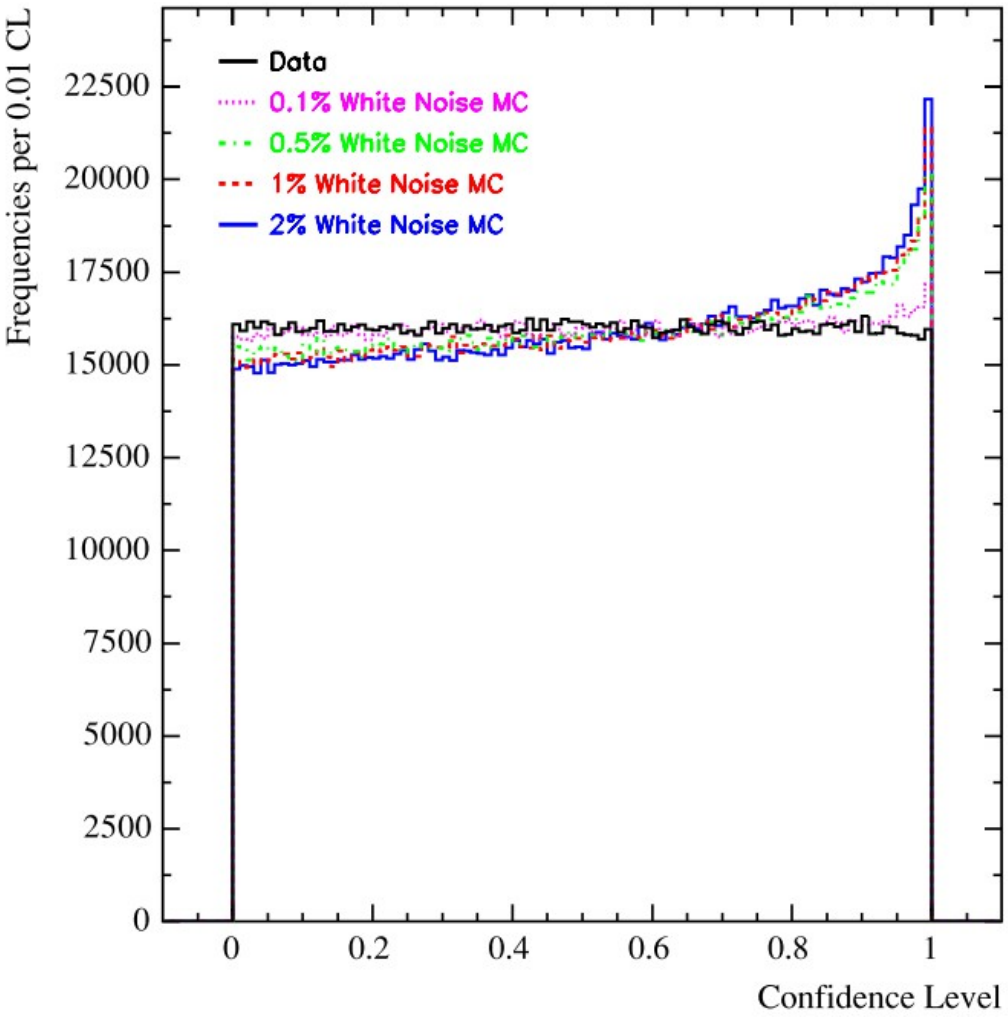
Keys to the analysis:

- Unbinned analysis
- Explicit account taken of when detector is live
- Trials factors accounted for
- Must scan with frequency resolution a few times smaller than 1/T, where T is the duration of the data set.

Broadband search

What if the signal is not sinusoidal?

A broadband search is possible by examining the distribution of confidence levels for all frequencies tested. Here SNO tested for white noise in the 10 minute – 1 day range.



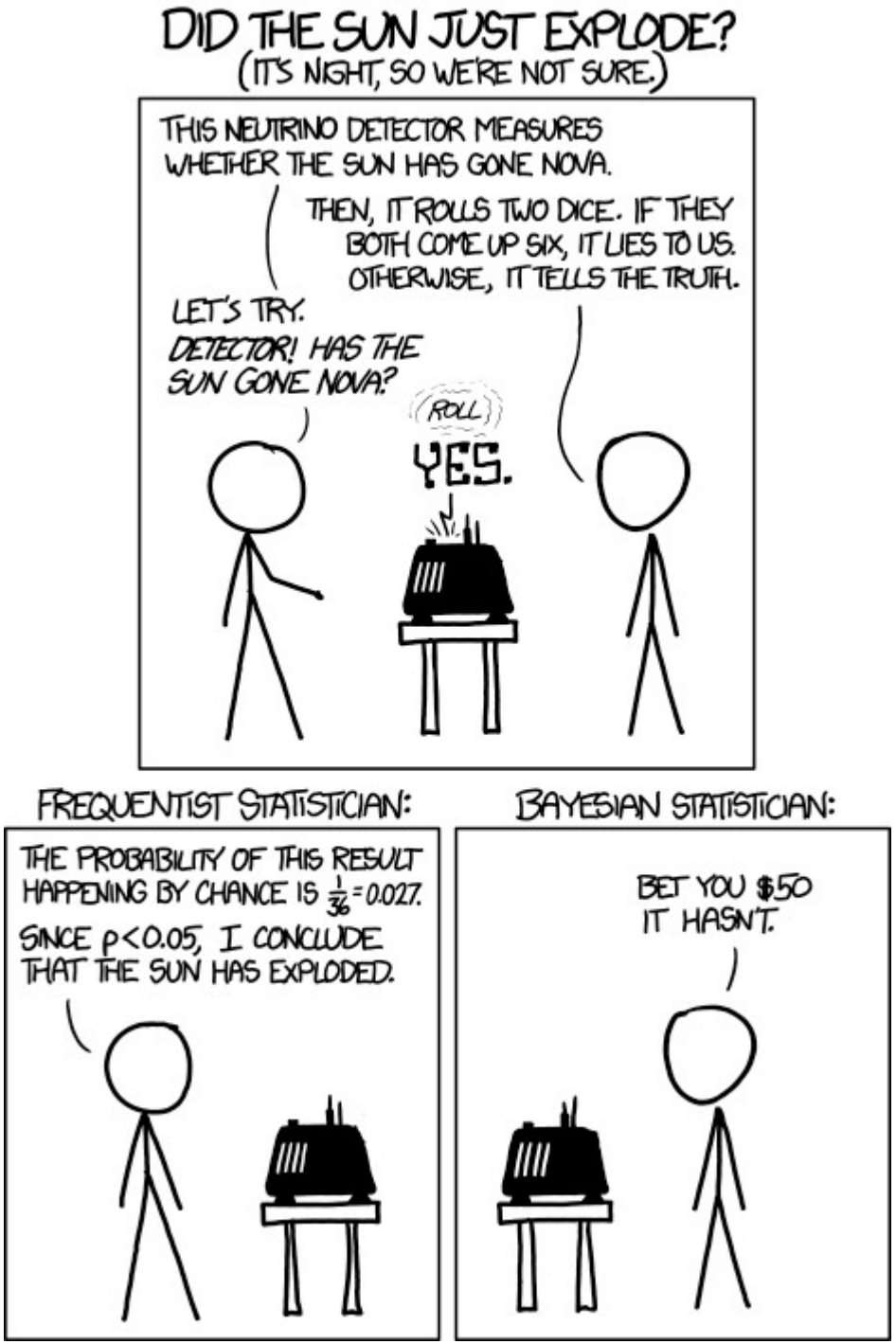
Non-sinusoidal time variation at specific frequencies can be tested with an *h*-test [see A&A 517, L9 (2010), DOI: 10.1051 / 0004-6361 / 201014362].

Time Variability Best Practices

- Avoid binning data if possible. If you must bin, use bins that are short compared to shortest period of interest.
- Account explicitly for livetime exposure, and beware of periodicities inherent in the exposure map.
- Use of toy MC to evaluate significance is mandatory.
- Be extremely wary of trials factors, including those associated with choice of analysis technique.

Conclusions

- Likelihood fits are ubiquitous in solar neutrino analyses, but computational subtleties must be considered when building PDFs from Monte Carlo.
- Time series analyses are a unique aspect of solar neutrinos for neutrino physics, but are statistically fraught. Proceed with caution!



https://xkcd.com/1132 ("Frequentists vs. Bayesians")