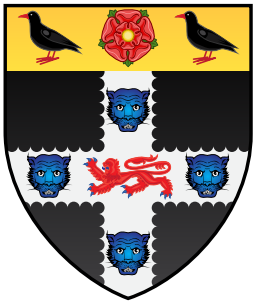
First Measurement of the Solar Neutrino Oscillation Parameters via Boron-8 Solar Neutrinos in SNO+



Daniel Cookman

Christ Church College

University of Oxford

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Trinity Term 2022

To my parents

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Chapter 1

The Theory of Neutrino Physics

Chapter 2

The SNO+ Detector

Chapter 3

Optical Scattering Theory

Chapter 4

Simulating SMELLIE events

4.0.1 Previous attempts at SMELLIE event simulation

Critical to extraction of scattering information from SMELLIE data is an accurate Monte Carlo (MC) simulation of the SMELLIE system. By modelling the laser light emission into the detector correctly, we can simulate how SMELLIE light will be impacted by changing scattering lengths in the detector. Because of the complexity of the optics of the optical fibres used to direct the laser light into the detector, a given SMELLIE event is simulated as a partially-collimated “flash" of visible photons emanating from the emission point of the fibre into the detector. This flash then requires a number of parameters to be correctly described. In particular, the fibre emission positions were recorded during the installation of the fibres. The wavelength and emission timing distributions of light pulses were taken from measurements of the laser heads by their manufacturers, or by colleague Jeff Lidgard in the case of the SuperK wavelength distribution. The pulse magnitude is defined as the mean number of photons simulated per event; on an event-by-event basis we sample Poisson fluctuations about that mean value. Determination of the pulse magnitude must be done on a subrun-by-subrun basis. Unlike scintillation light, light from SMELLIE is

not at all isotropic, and so we must specify some form of angular emission distribution. Determining and handling these angular emission distributions, also known as beam profiles, is the focus of this chapter.

Before we can determine the beam profiles, we must first decide how to specify them. Previous observations show that different fibres can have notably different beam profiles [2], so we let each fibre's beam profiles be unique. We assume for now that a given fibre's beam profile is stable over time, and independent of the wavelength of light fired. A straightforward, naïve approach to parameterising a beam profile would be as follows: specify some nominal fibre direction, corresponding to the direction light takes travelling from the fibre to the centre of the "beamspot" observed on the other side of the detector. Then, specify a 1D beam profile, corresponding to the probability density of firing a photon at a given polar angle α relative to the nominal direction. One might even assume this distribution is Gaussian. The distribution in azimuthal direction, ϕ, is assumed to be uniform.

This 1D beam profile approach was used initially for SMELLIE, and remains in use for the other ELLIE sub-systems within SNO+. However, when SMELLIE data was taken in the water-phase of the experiment, simulations using these beam profiles failed to match them well at all - see figure 4.1 for an example. Not only was the distribution in α not Gaussian, a distinct speckle-pattern can be observed within the beamspot that is not uniform in ϕ. This fact led to colleague Esther Turner building a SMELLIE generator that could handle 2D beam profiles: dependent on both α and ϕ. The distribution was stored as a map from each inward-pointing PMT in the detector to a relative intensity value. This was chosen because the beam profile shapes were calibrated from existing SMELLIE data — more on this in section 4.0.3.

This original 2D generator then sampled the beam profile via a rejection sampling approach, outlined as follows:

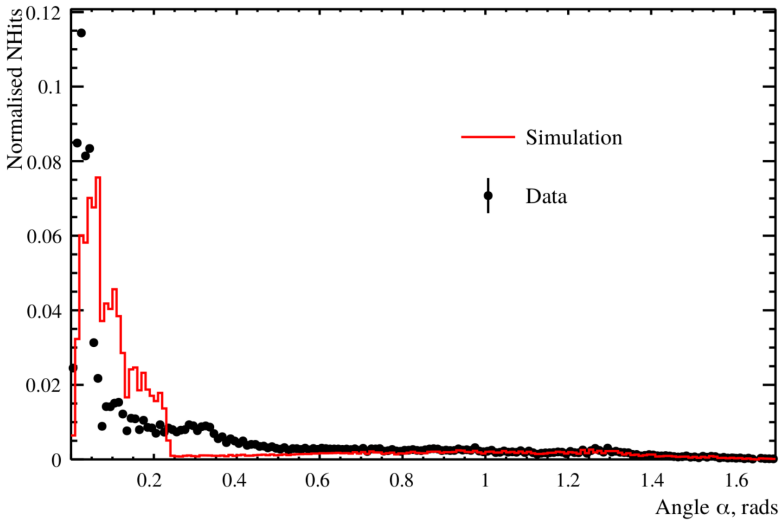


Fig. 4.1 Comparison between a simulation of one of the fibres, made from the 1D beam profile generator (red), with the associated data subrun that was used to create that beam profile (in black). For both MC and data, what is plotted is the PDF of observed PMT hits, as a function of the α angle. Poissonian errors have been added to the data points, but are too small to see. Clearly, this 1D generator does not replicate the observed beam profile correctly. Figure taken from [1].

1. Propose a test direction (α, ϕ), by generating ϕ uniformly in the interval $[0, 2\pi]$, and α according to some pre-determined Gaussian distribution, known as the Gaussian envelope.

2. Given this test direction, calculate where a line following this direction from the fibre of interest will hit the PSUP on the other side of the detector. Find the 3 closest PMTs to that point.

3. From those PMTs, obtain their relative intensity values from the beam profile mapping, and perform an interpolation based on how close each PMT is to the PSUP intersection point. This gives an interpolated relative intensity value for this test direction.

4. Because we are sampling using the angular coordinates (α, ϕ), differential area elements over this space of directions do not have the same size. We can correct

for this fact by multiplying our interpolated relative intensity by $\sin\alpha$, which corresponds to the Jacobian of the direction-space.

5. Calculate the value for the Gaussian envelope along this test direction.

6. Throw a random number uniformly between 0 and the Gaussian envelope value. If the random number is less than the interpolated intensity, then this test direction is accepted, and a photon is generated with that direction. Otherwise, we reject the direction and try the whole process again.

This generator certainly works, but has a key problem: efficiency. The 1D generator was able to generate a SMELLIE event (that is, to fully specify the starting parameters of all the photons emitted from a fibre) at a speed of $\sim 1\,\text{ms}$. However, the 2D generator specified here could take upwards of $\sim 50\,\text{s}$ *per event* to generate. Because a typical SMELLIE analysis requires simulating many millions of events, the CPU time taken to perform this quickly became unfeasible. Fixing this generator speed problem was a high priority for the SMELLIE analysis.

4.0.2 The new generator

On careful inspection of the existing 2D generator, the main reason for the slowness of the algorithm is the use of a rejection approach. Even with use of the Gaussian envelope, which was included to help with speed, the vast majority of proposed directions are never selected. Figure 4.2 shows a histogram of number of attempts per event it took for a valid direction to be chosen for a representative SMELLIE simulation. Moreover, the calculations needing to be done for every proposed direction are relatively complex, notably trying to find the 3 nearest PMTs to some point on the PSUP.

A new 2D generator was built with these thoughts in mind. Firstly, the rejection method would no longer be used, given its inefficiency. We would also endeavour to

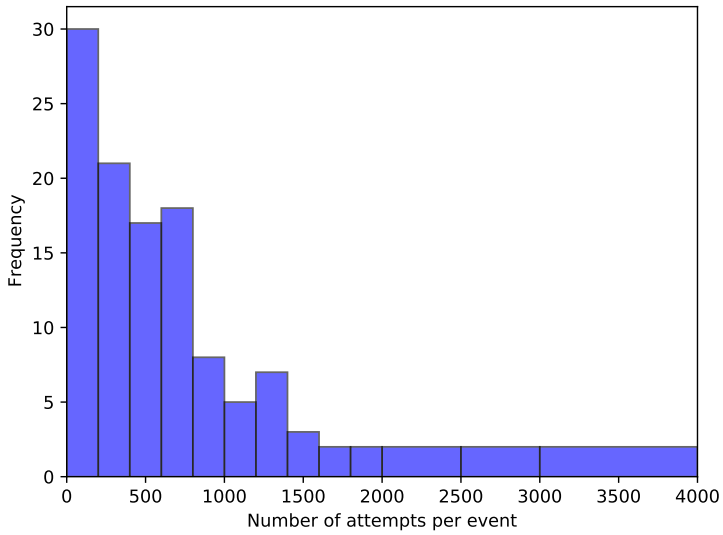
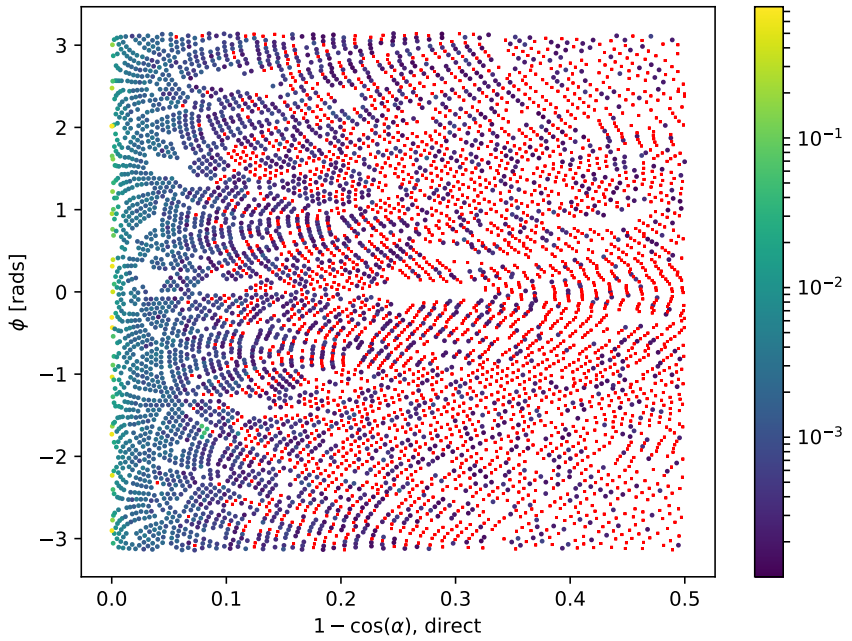


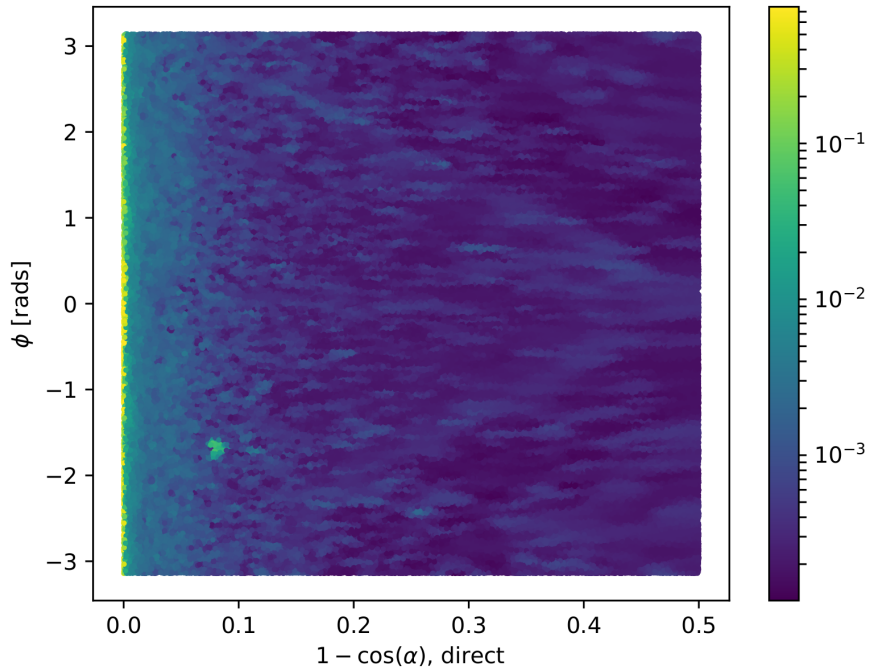
Fig. 4.2 Typical distribution of the number of attempts it takes for the existing 2D generator before the test direction gets accepted, per event.

try and "pre-calculate" as much as possible before run-time. Starting with the existing PMT relative intensity maps, we plot these in the 2D direction-space $(1 - \cos\alpha, \phi)$: see Figure 4.3a. In a toy-MC simulation, 500,000 directions are then thrown uniformly in this 2D space per fibre. For each direction, the same method of obtaining an interpolated intensity value from the nearest PMTs to the corresponding point on the PSUP as from the original 2D generator was performed, the only difference being that these calculations were done well before any actual SMELLIE simulation. Figure 4.3b shows the interpolated intensities obtained for one fibre.

Following this, the sampled intensities were then binned into a 2D histogram, where the bin value corresponds to the sum of all intensities for all directions found within this bin. Choosing a sensible binning procedure is important: too few bins, and necessary information about the shape of the beam is lost, whilst too many bins can oversample the data and capture statistical artefacts in the sampling process instead of just the beam profile. As a balance, 15 bins were chosen along the ϕ direction, and 60 in



(a) PMT relative intensity map



(b) Interpolated intensity map

Fig. 4.3 The first step in the new method for preparing the new generator. In (a), the relative intensities used for the existing beam profile of fibre labelled FS055 are shown for each PMT, the position on the plot indicating the location of that PMT in the fibre coordinates. The colour indicates the relative intensity; PMTs marked red have an intensity of zero. Figure (b) shows the result of throwing 500,000 directions uniformly over this 2D space, the intensity of each point given by interpolating the intensities of nearby PMTs.

$r = 1 - \cos\alpha$. This was chosen to ensure that a reasonable number of PMTs were located within each bin, lessening the impact of any statistical fluctuations. Although the bins in ϕ were chosen to have uniform width, this was decided to be not the case for the other axis, as there is far more important information near $r = 0$ (the beamspot). Instead, the width of the bins in r were calculated so that roughly the same total probability was contained in each r-strip. By consequence, bins near the beamspot typically are of significantly smaller size than ones much further out. This allows us to both capture any rapid changes in intensity near the beamspot, where this matters greatly, and smooth out the very-low intensities seen at larger polar angles. One of these histograms can be seen in Figure 4.4: the large change in bin widths as a function of r is clear. One can also see that near the beamspot notable dependence on the intensity as a function of ϕ. The mysterious "spot" at $r = 0.08$, well out of the beamspot, is an indication that the underlying beam profile data being used requires improvement: more on this in section 4.0.3.

The Cumulative Density Function (CDF) of this intensity histogram as a function of bin was then produced, where the bins were ordered through a raster-scan: scanning first over ϕ, and then r. The CDF was then normalised to 1 so that it was well-defined. It is this CDF object that is then loaded in and sampled from during event generation. To do this, an "inverse-CDF" approach was used, which has the major benefit over rejection sampling of always producing a valid direction for every sample made. The algorithm works as follows:

1. Throw a random number uniformly in $[0, 1]$.
2. Perform a binary search to find the bin that has the largest CDF value below this random number.
3. Look at the bin edges in ϕ of this selected bin: use linear interpolation of the random number to obtain a ϕ value located between these two ϕ-values.

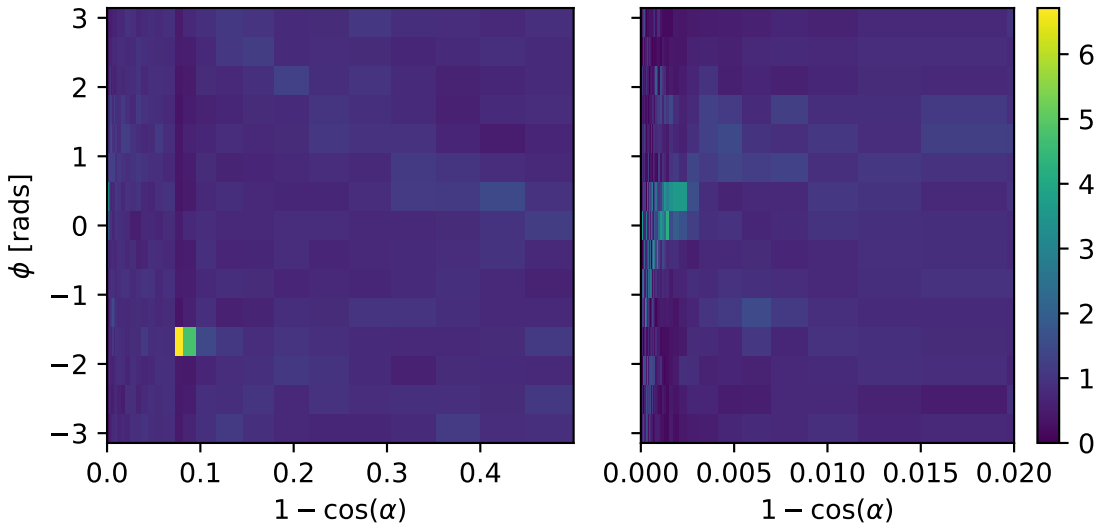


Fig. 4.4 Histogram of interpolated intensities within the 2D direction-space. The left view shows the full histogram; the right is a zoomed-in version near the beamspot. Unlike the binning in ϕ, the bin widths in r are not at all uniform. Instead, they have been determined such that the area summed over a given "strip" of bins of constant r will be the same.

4. Look at the selected bin's r-bin edges, and select a value of r by throwing a second random number uniformly between the two edges. Convert this r into a polar angle α.

5. The photon's direction is defined by the (α, ϕ) chosen by this process.

Because of the relative simplicity of this algorithm compared to the previous 2D generator, the speed improvement was very large: generation now took $\sim 1\,\text{ms}$ per SMELLIE event, a speed improvement of nearly 50,000. Event generation became as fast as it was when the 1D generator was being used. Furthermore, because of the approach taken, this major speed improvement comes at no sacrifice in accuracy. Figure 4.5 shows a comparison of the average number of photoelectrons (npe) per event per PMT between water-phase SMELLIE data and simulations with both the old and new 2D generator. One can see clearly that both generators are as accurate as one

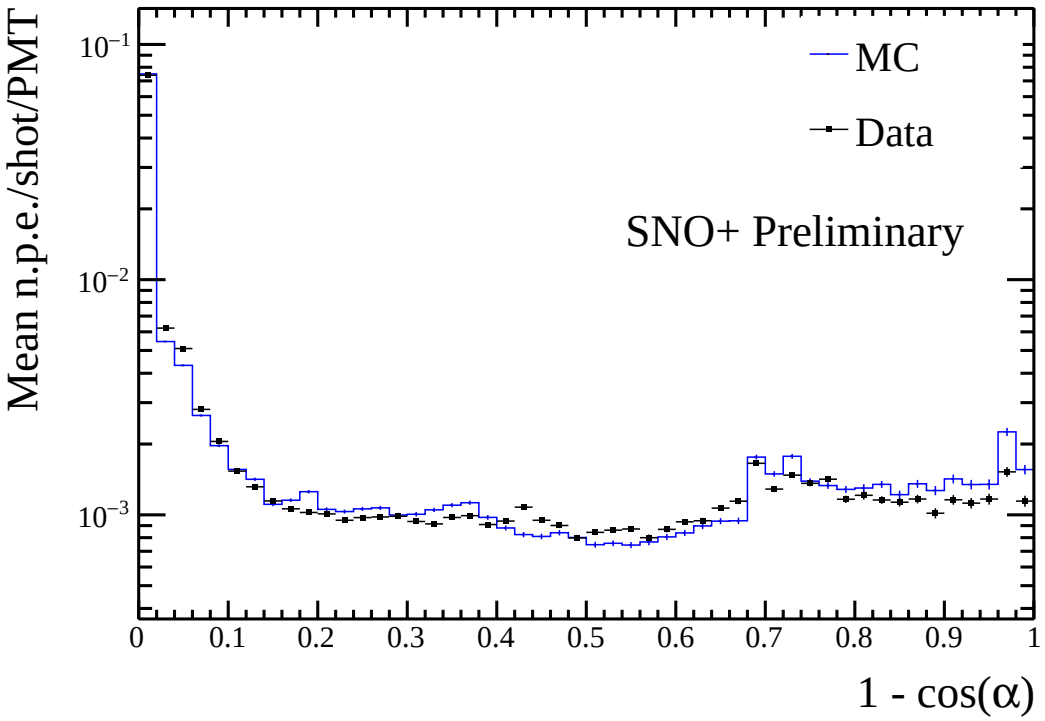


Fig. 4.5 Comparison of water-phase data to MC generated using both the old and new 2D beam profile generator approaches, with the updated beam profiles. Both versions of the generator are consistent with one another, but the new generator is many times faster.

another. Note that this plot uses the updated beam profiles as explained in the next section.

4.0.3 Improving the beam profiles

Even with the new 2D profile generator, a problem remains: the simulation fails to reasonably recreate data, and much of this appears to be because of the poor beam profile data being used. The curious "spot" for one of the fibres was already noted in the previous section that doesn't seem to be physical, and more broadly at large angles for all the fibres there are large swathes of PMTs with an intensity of zero, providing little useful information about the beam shape. It was shown in [1] that with the old

2D generator, the systematic uncertainty on the beam profiles was the dominant source of error in the main SMELLIE analysis. To help improve this situation, it was decided to update the existing beam profiles.

These old beam profiles were originally determined by looking at SMELLIE data taken during the water-phase. Specifically, a "medium"-intensity subrun with one of the lasers firing at a wavelength of 495 nm, was chosen for each fibre. "Medium"-intensity corresponds to firing the relevant laser at a set intensity determined during an earlier commissioning process, for which the maximum occupancy of PMT hits at that intensity, i.e. the proportion of hits per event, corresponded to roughly 80%. This value was chosen as it allowed for high statistics in a relatively short run-time, but not so intense that the occupancy of any given PMT in the beamspot was 100%. Because Rayleigh scattering is strongly-dependent on wavelength, the long wavelength of light was chosen so that impacts from this scattering were small in the data.

SNO+ PMTs are unable to distinguish the exact number of photoelectrons being generated. One is typically only able to know if a PMT has been triggered at all, by any number of photoelectrons. As a result, the occupancy of a PMT over a number of SMELLIE events, o, is a biased estimator of the mean number of photoelectrons generated, μ. Assuming the number of photoelectrons generated in a given event follows Poisson statistics, the probability of generating k photoelectrons is:

$$P\left(k|\mu\right) = \frac{\mu^k e^{-\mu}}{k!}. \tag{4.1}$$

The probability of observing a "hit" in a given PMT corresponds to generating at least one photoelectron:

$$P\left(\text{hit}|\mu\right) = P\left(k \geq 1|\mu\right) = 1 - P\left(k = 0|\mu\right) = 1 - e^{-\mu}, \tag{4.2}$$

Run Number	Run Type	Comments
114,018	All PQ lasers; SuperK laser in 400–500 nm range	Only PQ495 laser and SuperK at 495 nm is used
114,023	SuperK laser in 500–600 nm range	Part 1 of this wavelength range; crash occurred on last subrun, so that subrun is ignored
114,034	SuperK laser in 500–600 nm range	Part 2 of this wavelength range

Table 4.1 Water-phase runs used for new beam profiling.

which implies after rearrangement that one can determine the mean number of photo-electrons per event from the occupancy by:

$$\mu = \ln(1 - o)\,. \tag{4.3}$$

This is the reason why we want to avoid PMTs with occupancies of 100%: they preclude one's ability to convert into a value for μ by looking at occupancy alone. We call this conversion from occupancy into npe the "multi-hit correction". The impact of this correction is typically small for most PMTs, but can become very significant in a fibre's beamspot.

Once the npe mapping from data was obtained, a correction was then made for the detector's optics: even ignoring a fibre's beam profile, we still expect certain PMTs to be illuminated more than others because of e.g. reflections off the AV, or the solid angle subtended by the PMT bucket opening. For each fibre, a simulation was made where the beam profile was set as isotropic, and the corresponding npe mapping obtained: this map held information about the detector optics only. The beam profile mapping was then derived by simply dividing each fibre's npe mapping from data to its associated isotropic MC npe map. It is these maps that were first used in section 4.0.2.

4.0.4 Combining beam profile datasets

Fortunately, much more SMELLIE data was taken during the water-phase than was used for the original beam-profiling analysis. This additional data can be combined with that which was already used to far better constrain the beam profiles. In particular, given the existing assumption that scattering effects are minimal above wavelengths of $\sim$ 490 nm, all data taken with wavelengths above this can also be used. The specific runs (and associated comments about their specifics) are described in Table 4.1. Because high-intensity runs require a different analysis approach (PMTs with high occupancies must use charge, not occupancy, to estimate npe), for this analysis we only considered subruns that used low or medium intensity set-points.

For each subrun j of data per fibre, we look only at PMT hits for each PMT i that has been identified as "good" for that subrun[1], $i \in G_j$. G_j here represents the set of good PMTs in subrun j. In particular, a "good" PMT must have valid electronic and timing calibrations, be at high voltage and masked into the detector's trigger system for that subrun. In addition, an angular cut of $\alpha < 60°$ was made to remove PMTs that are well outside any reasonable beam direction. To isolates the hits arriving directly from the fibre without reflecting, scattering, or being noise, a time cut was also made. Because what matters is the time relative to emission from the fibre, and the expected time-of-flight from fibre to different PMTs varies, a quantity known as the time residual was used. Starting with the calibrated hit time of a given PMT relative to the event's trigger time, t_{hit}, the expected time-of-flight t_{TOF} from the fibre to the PMT was subtracted, estimated with the collaboration's "Light Path Calculator". Then, the emission time was also subtracted, t_{emm}, estimated by looking at the second-earliest value of $t_{hit} - t_{TOF}$ within the fibre's central beamspot, defined as the PMTs for which $\alpha < 3°$. It was found that a "loose" time residual cut of $t_{res} \in [-10, +12]$ns was

[1]Strictly speaking, a PMT's "goodness" is only determined on a run-by-run, not a subrun-by-subrun level, but this has no impact on the analysis.

sufficient to remove the vast majority of non-direct light with little signal sacrifice. In the situation where a subrun with intensity was very small, it would not regularly have at least two hits in the beamspot, and so the time residuals calculated would not be valid for many events. To avoid this situation, a cut was made on any subruns with mean intensities below 9 within their beamspot. This value was chosen as it would mean a 2σ fluctuation downwards of $2 \cdot \sqrt{9} = 2 \cdot 3 = 6$ npe would still have more than the 2 hits necessary for timing reconstruction. One fibre, FS207, has no data subruns that satisfy this condition, and as such will have to be dealt with separately. For the time being, this fibre was ignored.

Extracting the underlying beam profiles from these data required some careful thought, especially because different subruns could have different intensities. Considering a PMT i in subrun j, the mean number of photoelectrons generated per event in that PMT for that subrun, μ_{ij} can be decomposed as follows:

$$\mu_{ij} = I_j k_i = I_j b_i f_i. \tag{4.4}$$

I_j is the intensity of the subrun, i.e. the mean number of photons generated from the fibre in that subrun per event. k_i is the probability that a given photon generated at the fibre source ends up generating a photoelectron in PMT i. This itself can be further split into two components: b_i, the probability that a given photon at the fibre source points in the direction of PMT i; and f_i, the probability that a given correctly-pointed photon actually makes it to the PMT and successfully generates a photoelectron. It is b_i that is the actual beam profile we would like to measure.

Letting p_{ij} be the probability of observing a hit for a given event on a given PMT, the probability of observing m_{ij} hits out of N_j events in the subrun will be

binomially-distributed:

$$P(m_{ij}|\mu_{ij}) = L(\mu_{ij}|m_{ij}) = \binom{N_j}{m_{ij}} p_{ij}^{m_{ij}} (1-p_{ij})^{N_j - m_{ij}} = \binom{N_j}{m_{ij}} \left(1 - e^{-\mu_{ij}}\right)^{m_{ij}} e^{-\mu_{ij}(N_j - m_{ij})}. \tag{4.5}$$

Here we have used equation 4.2, and noted that this probability distribution in m can be re-framed as a likelihood function for the parameter μ_{ij}. Considering only a single subrun of data, the maximum likelihood estimate of the parameter μ_{ij} can be shown to be:

$$\langle \mu_{ij} \rangle = -\ln\left(1 - \frac{m_{ij}}{N_j}\right) = \ln\left(1 - o_{ij}\right) \qquad (m_{ij} \neq N_j), \tag{4.6}$$

where o_{ij} is just the occupancy of PMT i in subrun j. This is just the multi-hit correction formula seen in equation 4.3, which makes sense.

When looking at multiple subruns for the same fibre, the total likelihood function for a given PMT when considering all the data for a given fibre will be the product of the likelihoods from each dataset,

$$L\left(\{I_j\}, k_i | \{m_{ij}\}\right) = \prod_j L(I_j, k_i | m_{ij}) = \prod_j \binom{N_j}{m_{ij}} \left(1 - e^{-I_j k_i}\right)^{m_{ij}} e^{-I_j k_i (N_j - m_{ij})}. \tag{4.7}$$

This leads to a log-likelihood distribution of

$$\mathcal{L}\left(\{I_j\}, k_i | \{m_{ij}\}\right) = \sum_j \left[\ln\left({}^{N_j}C_{m_{ij}}\right) + m_{ij} \ln\left(1 - e^{-I_j k_i}\right) - I_j k_i \left(N_j - m_{ij}\right)\right]. \tag{4.8}$$

Formally, one could combine the likelihoods of all the PMTs together, and by looking at the maximum likelihood estimates for each of the parameters measure the parameter values this way. However, the set of equations one obtains through this approach quickly become analytically intractable, because the PMTs are coupled by the intensity values I_j. Even a direct numerical approach would be liable to fail: for a given fibre

there can be dozens of subruns, and many thousands of PMTs of relevance, so the dimensionality of the system of equations would be far too large.

Because of this, a different approach was taken. It is expected that in a subrun the total npe, summed over all good PMTs, should be proportional to the intensity value I_j. One must be careful about this construction — different subruns can have different sets of good PMTs, so two subruns with identical I_j values could have a larger summed npe merely because more PMTs were good in that subrun. To counter-act this effect, only PMTs that were classified as good in *all* subruns being analysed for that fibre would be used for the npe summation. In other words, we use data from PMT i for summing only if:

$$i \in \mathcal{I} = \bigcap_j G_j. \tag{4.9}$$

We can then define the summed npe for a given subrun as $S_j = \sum_{i \in \mathcal{I}} \mathrm{npe}_{ij}$, and assert that $I_j = cS_j$. By finding a value proportional to I_j, there is now enough information to maximise the log-likelihood $\mathcal{L}\left(k_i | \left\{m_{ij}\right\}, \left\{I_j\right\}\right)$ with respect to k_i for each PMT independently, and hence obtain estimates for these k_i parameters.

Of course, what is actually wanted are the underlying b_i values, not k_i. This is where isotropic simulations come in. For each run of data used, a matching isotropic MC was produced. As an example, a simulation for run 114,023 contained 200,000 events for each fibre using an isotropic beam profile, over the full wavelength range considered in this run, 500–600 nm, using the same run conditions as in data (which PMTs were at high voltage, etc.).

For each isotropic MC run, both I_j^{MC} and k_i^{MC} were calculated via the method described above. Because the simulations were isotropic, the underlying value for b_i was constant across all the PMTs, and so $ak_i^{MC} = f_i$. By doing some rearranging of

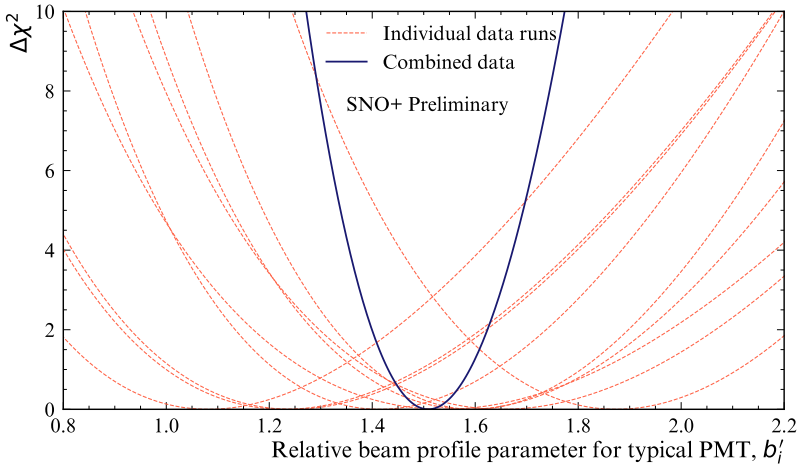


Fig. 4.6 Plot of $\Delta\chi^2 \simeq X_i$, twice the negative log-likelihood ratio, for both single subruns of a typical PMT, and when all relevant subruns are combined.

equation 4.4, we find that:

$$\mu_{ij} = I_j b_i f_i = c S_j b_i a k_i^{MC} = (acb_i) S_j k_i^{MC}. \tag{4.10}$$

As a result of this, given the set $\{S_j\}$ and k_i^{MC}, one can maximise the log-likelihood $\mathcal{L}$ with respect to $b_i' = acb_i$ numerically, to obtain the maximum likelihood estimate of b_i'. Because a and c were global constants of proportionality, they would become irrelevant as soon as the beam profile was normalised in the CDF-creation process outlined in 4.0.2.

Figure 4.6 shows the shape of this log-likelihood distribution for a particular PMT when considering fibre FS007's beam profile. One can see how individual subruns provide much more information when combined than if one looked at a single subrun alone.

Another benefit of using this log-likelihood approach is that the resulting distribution's shape can be used for uncertainty estimation. In almost all cases, Wilks Theorem [3] allows us to produce 1σ confidence intervals about the maximum likelihood

estimate for b'_i, $\langle b'_i \rangle$, because

$$X(b'_i) = -2\left[\mathcal{L}(b'_i) - \mathcal{L}(\langle b'_i \rangle)\right]$$

approximates a χ^2-distribution. As a result, the error bounds on our parameter estimate are given by when $X = 1$. The fact that the shape of X can be well-approximated by a quadratic in the region near $X = 0$ indicates the validity of Wilks' Theorem being used here.

Only a couple of exceptions to this approach of parameter estimation are possible. In the case where $m_{ij} = N_j$, i.e. a PMT has 100% occupancy, no maximum likelihood estimate exists: we need not worry about this, as subruns where this occurs have not been used. On the other end, however, there are some PMTs for certain fibres where after all subruns of data have been included, there remains no hits. In this scenario, one can show that the log-likelihood becomes linear in the beam profile parameter:

$$\mathcal{L}(b'_i | \{m_{ij} = 0\}) = b'_i k_i^{MC} \cdot \sum_j [I_j N_j]\,. \tag{4.11}$$

This scenario is very much reminiscent of rare-decay searches, and a similar approach can be used. A 1σ upper limit on the possible value for b'_i can be analytically-calculated to be:

$$b'_{i,ulim} = -\frac{k_i^{MC} \sum_j [I_j N_j]}{\ln\left[1 - \mathrm{erf}\left(1/\sqrt{2}\right)\right]}, \tag{4.12}$$

where $\mathrm{erf}(x)$ is the error function.

4.0.5 Results & Discussion

Figure 4.7 shows the impact of using additional subruns of data on a typical beam profile. One can clearly see the great reduction in the number of PMTs with no hits

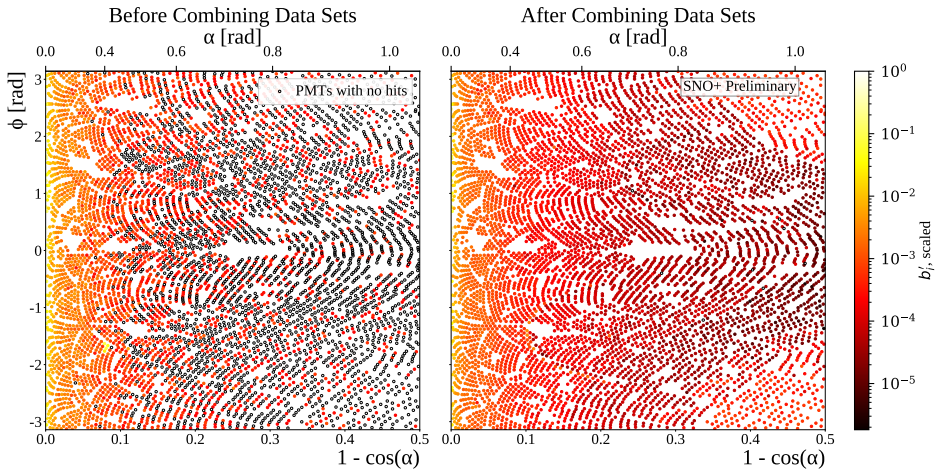


Fig. 4.7 Comparison between old and updated beam profiles for fibre FS055, after combining multiple data sets. Once again, the relative intensities (b'_i) for each PMT are given by the colour of each point, the position of each plotted in the 2D (r, ϕ)-space. The relative intensities have been both scaled here so that the largest value equals 1. Hollowed-out points are PMTs that, even after all relevant subruns have been combined, have no PMT hits.

in data. That many more data sets were included allowed for the major increase in dynamic range available for measuring these b'_i values. One can also note that by including additional data the curious spot that was seen in the old beam profile our at $r \approx 0.08$ has gone, further indicating that it was an artefact of that single data set.

Further details can be gathered from the interpolated intensity maps, one of which can be seen in figure 4.8. There are two curious stand-out features that can be seen here: firstly, there are multiple distinct parabolic arcs. These correspond to the shadows of the ropes that hold up/down the AV. More precisely, they are the mismodelling of those shadows — if the shadows were in the right place in the isotropic MC, then they would correctly cancel out any decreased intensity seen in the data of shadowed PMTs. These shadows could be mismodelled either because the positions of the ropes in the MC are in the wrong place, or the fibre's emission position is wrong. Note that any mismodelling of the fibre's nominal emission direction has no impact on this shadowing problem, as changing that direction merely causes a change of basis in the (r, ϕ)-space.

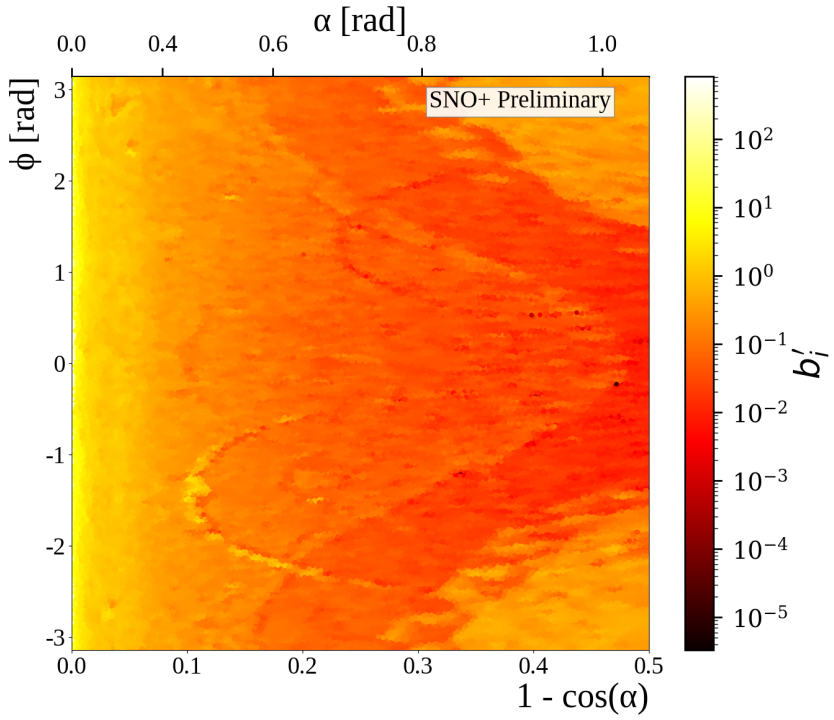


Fig. 4.8 Interpolated intensity map for the new updated beam profile of fibre FS055. The misalignment of rope shadows and AV effects, can both be seen.

The latter possibility of incorrect fibre positions are more likely, and in fact these arcs in the beam profiles could be used as an effective way to correct for this problem.

The second distinctive feature of this intensity map is the large band of lower intensity varying between $r \approx 0.2 - 0.5$, followed by larger intensity out at large r values. This feature comes from light reflecting off the AV surface, or internally-reflecting. The reason for this band's functional dependence on ϕ is that this particular fibre, FS055, has a nominal fibre direction $\sim 10°$ from pointing radially-towards the detector's centre. This feature appears in the updated beam profiles of all fibres, but its shape depends on the particular fibre's direction — for fibres pointing directly towards the detector's centre, there is little ϕ-dependence observed. Like the ropes, this feature must come from some form of mismodelling of the optics of the AV. A de-facto shadowing of PMTs in line with tangents from the AV surface which intersect

the fibre position is to be expected. One also expects PMTs at polar angles larger than this to have their observed intensities boosted from reflected light off the AV. However, the discontinuities seen in the beam profiles indicate that for whatever reason this effect has been over-emphasised in the simulation.

There is a further phenomenon that can be seen, by comparing beam profile values obtained from a single subrun to the updated combined beam profile. This can be done by calculating the residuals corresponding to the single subrun, relative to the combined data set. The residual is negative if the combined data sets have a b'_i below the equivalent for a given single subrun; that is, the combined model underestimates this subrun for that PMT.

This information was plotted for two different subruns from the same fibre, seen in figure 4.9. One subrun was the same one used by Esther Turner for the original 2D beam profiling, with a wavelength of 495 nm; the latter was at the longer wavelength of 595 nm. For both subruns, most PMTs are seen to have intensities well-modelled by the combined model. However, their appears to be a significant amount of mismodelling within the beamspot. There also appears to be some systematic shift between data and model at somewhat larger polar angles. Moreover, this mismodelling seems not to be merely random, but a function of wavelength: at shorter wavelengths the beamspot tends towards being overestimated and then underestimated at larger values of α. At longer wavelengths, the beamspot becomes underestimated, with larger angles getting overestimated. This indicates that there appears to be a wavelength-dependence on the beam profiles, contradicting one of the main assumptions which we used to combine the water-phase data in the first place! All three of these features — rope shadows, AV reflections, and wavelength dependence — add systematic uncertainty to the beam profiles, beyond the statistical uncertainty as measured by the width of the likelihood

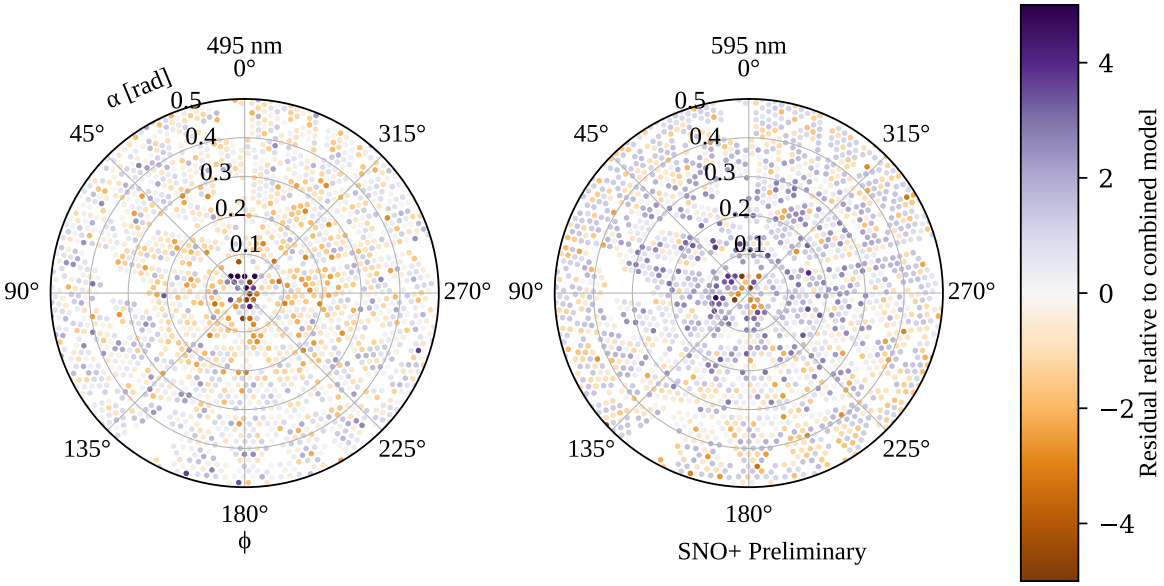


Fig. 4.9 Residuals from subruns at two different wavelengths, both compared to the combined beam profile model for fibre FS055. A negative sign, and hence bluer colours, indicate that the combined model underestimates the observed intensity for that particular subrun. Values with a magnitude beyond 5 are shown capped at this maximal value for the purposes of this plot. These PMTs are plotted in the polar fibre coordinates (α, ϕ).

distribution. Certainly if one wanted to further improve the uncertainties in the beam profiles, tackling these challenges would be key.

Chapter 5

Solar Oscillation Analysis

Driving out into the Sun
Let the ultraviolet cover me up
Looking for a Creation Myth
Ended up with a pair of black lips

This is the End
Phoebe Bridgers

Measuring the "solar" neutrino oscillation parameters Δm^2_{21} and θ_{12} is one of the principal aims of the SNO+ detector during the scintillator-phase. There are, in fact, two complementary methods of measuring these parameters: the oscillations of anti-neutrinos from terrestrial nuclear reactors, and the oscillations of neutrinos from the Sun.

This chapter focuses on the latter approach, using ^{8}B neutrinos coming from the Sun to measure the solar oscillation parameters. An initial background-free study was performed by Javi Caravaca [], which demonstrated that it was indeed possible to make such a measurement in the detector. The work in this chapter builds on substantially from that analysis. This chapter also draws on the associated reactor anti-neutrino

analysis built by Iwan Morton-Blake [], and more broadly from the general techniques used in the $0\nu\beta\beta$ analysis of Tereza Kroupova [] and Jack Dunger [].

This chapter begins by explaining how it is possible to measure the solar oscillation parameters via ^{8}B events. Then, the framework used to perform the analysis is then explained: that of a *Bayesian Analysis using Markov Chain Monte Carlo techniques*. After the method has been described, the dataset upon which the analysis is performed is then introduced. The results and associated validation are then given. Given these results, a projection is then made for the expected sensitivity to θ_{12} as a function of livetime.

5.1 Analysis Methodology

5.1.1 Observational Principle

How can we measure neutrino oscillation parameters via solar neutrinos in the SNO+ detector? As discussed in , it is possible to detect all flavours of neutrino through elastic scattering with electrons in the detector. If this interaction was purely neutral-current, then there would be no way of telling the flavour-state of an interacting neutrino. However, electron neutrinos are able to interact through an additional charged-current mode. This modifies the cross-section for electron neutrinos, and means that as the survival probability for electron neutrinos generated from the Sun, P_{ee}, is modified, the interaction probability of neutrinos with the detector will also.

Of course, we do not directly measure neutrino energies in the detector — only the associated scattered electron. If there were no correlation between the observed electron energy and its associated neutrino, then the only effect of neutrino oscillations would be to change the overall observed rate of events due to this process. There would be no change in the shape of the event's energy spectrum, even though neutrino

Fig. 5.1

Fig. 5.2

oscillations are a function of neutrino energy. Fortunately, there is some dependence of the neutrino's energy on that of the scattered electron. This dependence can be seen in Fig. 5.1 for ^{8}B electron neutrinos interacting in SNO+. As can be seen, the dependence is weak, and comes mostly from basic energy conservation: If one observes a 10 MeV electron event in the detector, it can't reasonably have come from a 5 MeV neutrino.

In Fig. 5.2 we can see the impact each physical process has on the energy spectrum that we eventually observe. We start with a broad energy distribution of ^{8}B electron neutrinos generated in the Sun. These neutrinos then oscillate their flavour state as they propagate to the detector, in an energy-dependent manner. When a (tiny) fraction of these neutrinos interact with the electrons in our detector, there is both an energy- and flavour-dependence on the cross-section. The scattered electrons gain a kinetic energy with some mild dependence on the inciting neutrino's energy, which is then measured by the detector to within some energy resolution.

Let us now consider the dependence of P_{ee} on the individual neutrino oscillation parameters. Recall from Eq. 5.1.1 that, after considering matter-induced oscillations due to neutrinos passing through the Sun and possibly the Earth, $P_{ee} = P_{ee}\left(\tan 2\theta_{12}^M, \sin\theta_{13}^M, \Delta m^2_{21,M}\right) = P_{ee}\left(\theta_{12}, \theta_{13}, \Delta m^2_{12}, \Delta m^2_{13}\right)$. Fig. 5.3 shows the dependence of each of these four oscillation parameters on $P_{ee}(E)$. We can see that in reality only the two parameters Δm^2_{21} and θ_{12} have a substantial impact on $P_{ee}(E)$ and hence the observed electron energy spectrum. Because of this, for this analysis we will only

Fig. 5.3

Fig. 5.4

ever vary these two oscillation parameters, and keep θ_{13} and Δm^2_{13} at their current global fit values[1] of $\sin^2\theta_{13} = 0.0222$ and $\Delta m^2_{13} = +2.515 \times 10^{-3}\,\text{eV}^2$ [].

5.1.2 Background Processes

Sadly, elastically-scattered electrons from ^{8}B neutrinos are not the only events we see in the SNO+ detector during the scintillator phase. There are a number of background processes that our signal must compete against. Below a reconstructed energy of $\sim 2\,\text{MeV}$, it is known that various backgrounds completely swamp any possible ^{8}B signal, and so for this analysis we only consider processes that can generate reconstructed energies of at least $E_{\text{min}} = 2\,\text{MeV}$. The following subsections explain each of these backgrounds, as well as methods that have been used to mitigate them as much as possible.

Internal Uranium- and Thorium-Chain Backgrounds

Although every effort has been made to make the scintillator cocktail that fills SNO+ to be as radio-pure as possible, there inevitably remain trace amounts of the radioactive isotopes that derive from the decay chains of the ^{238}U and ^{232}Th isotopes. Fig. 5.4 shows these two decay chains. Fortunately, only a fraction of the radioactive isotopes in these chains actually are capable of generating events in the detector with energies above E_{min}: these have been highlighted in Fig. 5.4 in gold.

[1] We use the global fit results excluding Super-Kamiokande's atmospheric data, and assuming normal ordering of the neutrino mass hierarchy. This choice has a tiny impact on the magnitudes of these two fixed parameters, the main impact being the sign of Δm^2_{13}.

Of particular note are the decays of ^{212}Bi and ^{214}Bi. Both are capable of either α- or β-decays to Tl or Po isotopes, respectively. For the former, it is the subsequent β-decay of the Tl that can have a reconstructed energy above E_{min}. For the latter, the Bi decay is the part of the pair of decays that can lie above E_{min}. Although the α-decays here certainly have Q-values well above 2 MeV, the liquid scintillator quenches the observed energy well below E_{min}. The so-called "Bi–Po" decays are particularly special because the lifetimes of ^{212}Po and ^{214}Po are 300 ns and 164 µs, respectively, which are short enough to allow for highly-effective coincidence tagging.

There are two classes of Bi–Po event in the detector: "out-of-window" events for which the Bi and Po occur in separate event windows, and "in-window" events whereby the Bi and Po occur within the same event window. These lead to two distinct strategies for tagging these kinds of events. For out-of-window Bi–Pos, we look for a delayed coincidence of two events. Using the tagging algorithm suggested in [] as a starting point, the chosen procedure was as follows. There must be two events that trigger the detector within 4 µs of one another, and both have a valid `scintFit` position reconstruction within 2 m of one another. The delayed candidate event must also have at least 100 cleaned PMT hits. This very broad coincidence tagging procedure was designed to ensure that the cut was as *efficient* in tagging (and hence, rejecting) Bi–Pos as possible, whilst negligibly impacting the solar signal. This is in contrast to the cuts chosen by Rafael Hunt-Stokes in [], which try and obtain a highly *pure* sample of Bi–Po tags.

Of course, the above delayed coincidence procedure cannot catch any of the in-window Bi–Po events. For these, we use a different approach. Because two decays happened in the same event window, we expect to see two distinct peaks in the event's time residual spectrum. In order to look for this event topology, a likelihood-ratio classifier was run over events, first developed by Eric Marzec [] and re-coordinated for

Fig. 5.5

the $2.2\,\mathrm{g\,l^{-1}}$ LABPPO scintillator optics by Ziping Ye []. This classifier calculates the likelihood ratio between the null hypothesis of a $0\nu\beta\beta$ event (a proxy in this analysis for single-site events such as our ^{8}B signal) and the alternative hypothesis of an in-window Bi–Po event. The more negative the value of the result, `alphabeta212`, the greater the evidence there is for rejecting the null hypothesis of a single-site event. Events with `alphabeta212` < 0 were then rejected.

Combining both out-of-window and in-window Bi–Po tagging, the impact on ^{212}Bi–Po, ^{214}Bi–Po, and ^{8}B ν_e events can be seen in Fig. 5.5. We consider here only events that pass all other cuts used in this analysis: see section 5.1.1 for the specifics of the cuts used. Because of the different lifetimes of the decays, ^{214}Bi–Po decays predominantly fall out-of-window whilst ^{212}Bi–Po events are typically in-window. This explains why the out-of-window tagging is substantially better at cutting ^{214}Bi–Po decays, whereas the in-window tagging far better tags ^{212}Bi–Po decays. Overall, within the analysis region of interest (ROI), the two combined cuts are able to tag TODO% of ^{214}Bi–Po triggered events, TODO% of ^{212}Bi–Po triggered events, whilst retaining TODO% of ^{8}B ν_e signal events.

(α, n) Reactions

The impact of ^{238}U- and^{232}Th-chain isotopes does not simply at their direct decays. It is possible for the αs generated during these decays to undergo their own interactions with nuclei in the detector. Within the organic scintillator of SNO+, the dominant interaction of this type is when an α collides with a ^{13}C nucleus, emitting a neutron: $\alpha + {}^{13}\mathrm{C} \longrightarrow {}^{16}\mathrm{O} + \mathrm{n}$. This is known as an (α, n) reaction.

The topology of this reaction in the detector is a delayed coincidence, as shown in Fig. 5.6. For the prompt signal, there is the light emitted from the α just before,

Fig. 5.6

Fig. 5.7

and the n just after the (α, n). The neutron then thermalises and gets captured by another nucleus — usually hydrogen in SNO+ — which creates an excited state that then eventually decays, creating a γ that creates the delayed signal in the detector [].

As can be seen in Fig. 5.7, (α, n) interactions can lead to events reconstructed at a wide variety of energies, which could be an issue for this analysis. However, because they are delayed coincidence events with a typical decay time of $\sim 100\,\text{ns}$, the aforementioned out-of-window Bi–Po tagging algorithm also efficiently tags (α, n) events. Looking again at Fig. 5.7, simply by using the out-of-window Bi–Po tagger without any further modifications TODO% of events in the ROI are cut.

External Backgrounds

All materials within the SNO+ detector are radioactive, not just the liquid scintillator cocktail. This includes the acrylic, ropes, external water, and PMTs. These components have had their radiopurity "assayed" (that is, measured) throughout the detector's lifetime, often back to the construction of the original SNO detector itself. The materials other than the liquid scintillator are known to have far higher background levels, especially in the important ^{238}U- and ^{232}Th-chain backgrounds []. To distinguish between the inherent backgrounds within the scintillator, and the backgrounds from materials at larger radii, we use the terminology "internal" and "external", respectively.

Although there are numerous external backgrounds, with a suitably accurate and precise position reconstruction algorithm they can be suitably handled. The simplest approach is with a so-called "fiducial volume" (FV) cut: just throw out all events that reconstruct beyond some radius. The only external background events that will

Fig. 5.8

reach within the FV are those that have reconstructed very poorly, or have some long-distance radiation that manages to deposit radiation close to the centre of the AV. Because α and β radiation can only travel short distances through the detector, it is only γ radiation that can realistically travel far enough into the detector to be able to reconstruct anywhere near the centre. Moreover, the intensity of this γ radiation attenuates exponentially towards the centre of the detector, meaning only a tiny fraction of the total number of external events reconstruct within a 3.5 m FV, say. This strong radial-dependence can be seen in Fig. 5.8.

What this figure also demonstrates is that our solar signal has a completely different radial dependence to these backgrounds. As a result, if one considers not just the energy of events but also their reconstructed radius, then it is possible to get an additional handle on the external backgrounds. The FV cut can then be pushed further out to larger radii, allowing one to gain more signal statistics.

Work by Tereza Kroupova [] allows for additional means of distinguishing external backgrounds from the solar signal. The underlying assumption in the reconstruction of SNO+ events is that there was an electron at a single point, which is entirely valid for ^{8}B elastic scattering events. However, external backgrounds can fail this assumption in two ways. Firstly, these radioactive decays often generate γ radiation in addition to the main α/β particle, which creates a multi-site event. Because the `scintFit` position reconstruction algorithm is not prepared for a distribution of energy depositions in the scintillator, the t_{res} distribution will broaden. This allows an event classifier to be built that distinguishes between the t_{res} distributions of single-site events and externals, known as the "external background timing classifier". Secondly, because external backgrounds that do reconstruct close to the centre of the detector typically have a γ that travelled a long distance towards the centre of the detector, we expect the earliest

Fig. 5.9

light that hits the PMTs to arrive most often along the direction of the reconstructed position vector. A distribution of PMT hits for a given event as a function of their angular distribution relative to the direction of position reconstruction can be built, and compared to the expected distributions for single-site and external background events. This is known as the "external background topological classifier". Much like the classifier described in Section 5.1.2, the single-site events used for comparison were $0\nu\beta\beta$ events, but these have a similar single-site structure to the solar signal of interest in this analysis.

Fig. 5.9 shows the correlation between the two classifier results for both a typical external background, and ^{8}B ν_e events.

Cosmogenic Isotopes

The final source of background events are radioactive isotopes that form via collisions of cosmic rays with atomic nuclei, known as cosmogenic isotopes. Most of these isotopes are short-lived [], with lifetimes $\mathcal{O}(1\,\mathrm{s})$. Fortunately, the depth of SNO+ means that our rate of cosmic ray muons interacting with the detector is only 3 an hour []. Because the rate is so low relative to other experiments, relatively straightforward approaches to tagging and removing cosmic ray muons and their cosmogenic followers can be utilised without substantial loss of livetime. Events are tagged as a cosmic ray muon if they create a sufficient number of hits in the outward-looking PMTs above background levels, as well as many hits within the detector itself. The details of this tagging for the scintillator-fill were put in place by Lorna Nolan [], modifying the existing algorithms used in the water-phase [] and in SNO []. After a tagged cosmic muon event, all events for the next 20 s are then vetoed as a means of rejecting followers. This simple cut is enough to remove the vast majority of expected cosmogenic events

in the scintillator-phase. The expected impact on loss of livetime, and hence quantity of signal events, is 3 lots of 20 s vetoes an hour, that is to say $1/60^{\text{th}}$ of the signal is lost via this cut.

There is one cosmogenic isotope with a long-enough half-life that even the 20 s muon follower veto is not sufficient to remove all events. This is ^{11}C, which β^+-decays to stable ^{11}B with a half-life of 20.4 min. The maximum possible energy deposited in this decay is 1.982 MeV [], just below E_{min}, so only a small fraction of ^{11}C events end up in the ROI: the ones with very high energies that get their energy reconstruction falsely-inflated by some amount. As a result, this background is expected to be very much sub-dominant to all other backgrounds in this analysis. Because this background is important to consider for some other analyses, a triple-coincidence tagging algorithm is being built by Katharine Dixon [], but not used for this analysis currently.

5.1.3 The Log-likelihood Test Statistic

At the highest level, this analysis involves taking the data observed in the scintillator-fill after applying a certain set of cuts, along with simulated PDFs for all processes believed to build up the observed data with those same cuts applied, and then attempting to fit the combined energy and radial distributions of the MC to that of the data. Given a set of PDFs, to try and match the distribution of observables in data we can modify a number of parameters. These consist of the normalisations of each PDF (i.e. the total number of events observed due to that process), and any systematic parameters that could modify the shapes of these distributions. For this analysis, the neutrino oscillation parameters act as *de facto* systematic parameters, as they modify the shape of the ^{8}B PDFs. Of course, unlike usual systematics the oscillation parameters are what we are actively trying to measure, instead of being a nuisance.

In order to perform a fit to data in this way, we must first answer a set of questions:

1. Which signal and background processes must we consider?
2. In addition to their normalisations, are there any further parameters necessary to specify the distributions of the PDFs for each of the processes? Systematics and oscillation parameters are good examples.
3. What is our test statistic?
4. What algorithm do we use to try and find the best-fit result?
5. How do we measure uncertainties on these best-fit values for each parameter?

In section 5.1.2, question 1 was answered for this analysis. We now give the answer to question 3; all other questions on this list will be answered shortly.

The test statistic used for this analysis is the *binned extended log-likelihood.* Once the data and MC PDFs have been binned in both the observables of interest, it is assumed that the expected number of events in a given bin j is governed by a Poisson distribution:

$$P_j\left(n_j|\lambda_j\right) = \frac{\lambda_j^{n_j} e^{-\lambda_j}}{n_j!}, \tag{5.1}$$

where $P_j\left(n_j|\lambda_j\right)$ is the probability of observing n_j events in bin j, given an expectation of λ_j events in total from signal and background processes in that bin. This λ_j can be decomposed into each of the expected rates for each process, i:

$$\lambda_j = \sum_{i=1}^{N_{\text{PDFs}}} \mathcal{N}_i P_{ij}\left(\boldsymbol{\theta}\right), \tag{5.2}$$

where N_{PDF} is the number of PDFs being considered in the analysis, $\mathcal{N}_i$ is the normalisation parameter of the i^{th} PDF, and $P_{ij}\left(\boldsymbol{\theta}\right)$ is the probability of observing an event of process type j in bin i, assuming a set of non-normalisation parameters $\boldsymbol{\theta}$. By combining the probabilities of all the bins together, and also adding the possibility of constraining any of the normalisation or systematic parameters $\boldsymbol{\mathcal{N}}$ or $\boldsymbol{\theta}$, the total

probability for a given set of processes assuming these parameters to give rise to the data seen is:

$$P\left(\boldsymbol{n}|\boldsymbol{\mathcal{N}},\boldsymbol{\theta}\right) = L\left(\boldsymbol{\mathcal{N}},\boldsymbol{\theta}|\boldsymbol{n}\right) = \prod_{j=1}^{N_{\text{bins}}} \frac{\left[\sum_{i=1}^{N_{\text{PDFs}}} \mathcal{N}_i P_{ij}\left(\boldsymbol{\theta}\right)\right]^{n_j} e^{-\sum_{i=1}^{N_{\text{PDFs}}} \mathcal{N}_i P_{ij}(\boldsymbol{\theta})}}{n_j!} \cdot \prod_{k=1}^{N_{\mathcal{N}'}} \frac{e^{-\frac{\left(\mathcal{N}'_k - \mathcal{N}_k^{\text{nom}}\right)^2}{2(\sigma_k^{\mathcal{N}})^2}}}{\sqrt{2\pi\sigma_k^{\mathcal{N}}}} \cdot \prod_{\ell=1}^{N_{\theta'}} \frac{e^{-\frac{\left(\theta'_\ell - \theta_\ell^{\text{nom}}\right)^2}{2(\sigma_\ell^{\theta})^2}}}{\sqrt{2\pi\sigma_\ell^{\theta}}}. \tag{5.3}$$

Here, N_{bins} is the total number of bins being considered in the analysis. Gaussian constraints [2] on a subset of the normalisations have been included, $\{\mathcal{N}'\}$, of which there are $N_{\mathcal{N}'}$ in total. For each of these normalisations, indexed by k, there is an associated nominal value $\mathcal{N}_k^{\text{nom}}$ and width $\sigma_k^{\mathcal{N}}$. There are similar constraints on a subset of the systematic parameters, with similarly-named variables.

This probability can be re-framed as the likelihood of the vectors of parameters $\boldsymbol{\mathcal{N}}$ and $\boldsymbol{\theta}$ given the vector of number of events in each bin, $\boldsymbol{n}$: $L\left(\boldsymbol{\mathcal{N}},\boldsymbol{\theta}|\boldsymbol{n}\right)$. It is rare to see the likelihood as-is, instead, for computational purposes the log-likelihood is used instead, $\mathcal{L}\left(\boldsymbol{\mathcal{N}},\boldsymbol{\theta}|\boldsymbol{n}\right) := \ln L\left(\boldsymbol{\mathcal{N}},\boldsymbol{\theta}|\boldsymbol{n}\right)$. We can then get to the formula actually used for this analysis:

$$\mathcal{L}\left(\boldsymbol{\mathcal{N}},\boldsymbol{\theta}|\boldsymbol{n}\right) = -\sum_{i=1}^{N_{\text{PDFs}}} \mathcal{N}_i + \sum_{j=1}^{N_{\text{bins}}} n_j \ln\left(\sum_{i=1}^{N_{\text{PDFs}}} \mathcal{N}_i P_{ij}\left(\boldsymbol{\theta}\right)\right) - \sum_{k=1}^{N_{\mathcal{N}'}} \frac{(\mathcal{N}'_k - \mathcal{N}_k^{\text{nom}})^2}{2(\sigma_k^{\mathcal{N}})^2} - \sum_{\ell=1}^{N_{\theta'}} \frac{(\theta'_\ell - \theta_\ell^{\text{nom}})^2}{2(\sigma_\ell^{\theta})^2}. \tag{5.4}$$

5.1.4 The Bayesian Statistical Approach & MCMC

There are two main schools of statistical inference, “Frequentist” and “Bayesian”. In the former, probabilities describe the fraction of times a situation can be found within

[2] As we shall see in section 5.1.1, because this analysis uses a Bayesian statistical approach, it is more appropriate to talk of these constraints as non-uniform “priors” for the parameters.

the whole ensemble of possible worlds. For the latter, we care not about an ensemble of worlds but instead our degree of belief in this current one. We update our beliefs as we acquire knowledge of the world through Bayes' Theorem:

$$P(\boldsymbol{\mu}|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{\mu}|\boldsymbol{x})\, P(\boldsymbol{\mu})}{P(\boldsymbol{x})}. \tag{5.5}$$

Here, $\boldsymbol{\mu}$ is the set of parameters that model our system, $P(\boldsymbol{\mu})$ is our *prior* (pre-existing) distribution for those model parameters, and $\boldsymbol{x}$ is the data taken in our experiment. The updated, *posterior* distribution $P(\boldsymbol{\mu}|\boldsymbol{x})$ is then the prior multiplied by the likelihood of parameters $\boldsymbol{\mu}$ given observations $\boldsymbol{x}$, $L(\boldsymbol{\mu}|\boldsymbol{x})$, and divided by the total probability $P(\boldsymbol{x})$ of observing $\boldsymbol{x}$ under any circumstance.

Both approaches to statistics are widely-used in statistical analysis, in both particle physics and beyond. The Bayesian approach was used for this analysis, as it was believed that this helps keep transparent what assumptions are being made in the analysis. More precisely, the suggestions made by Biller & Oser in [] are followed: for parameters that do not have some pre-existing constraint, a flat prior is used. For the bulk of this analysis, uniform priors are assumed on the neutrino oscillation parameters Δm^2_{21} and θ_{12}, as the magnitudes of these parameters are now well-established. There is no major reason to prefer a flat prior in θ_{12} as opposed to $\tan^2\theta_{12}$ or $\sin^2\theta_{12}$; we will see the impact of these different choices of prior in section 5.1.1.

Now, if one is able to determine the overall posterior distribution, then it is possible to derive best-fit values with uncertainties for all parameters in the fit. This is done by "marginalising" the posterior distribution, i.e. integrating over all parameters other than the one of interest. A sensible best-fit value is then the modal marginalised posterior density, the highest value in the marginalised distribution. The uncertainty on this value is derived from the spread of the marginalised posterior, by the calculation of the 1σ Credible Interval (CI): this is the set of values for a given parameter which

has a total posterior probability of 68.3%, and contain the best-fit value. There are in fact an infinite number of CIs that satisfy this property; for this analysis, the values are chosen in decreasing order of marginalised posterior probability density.

Of course, all of this assumes that one can accurately determine the posterior density distribution. Whilst the likelihood and prior distribution are straightforward enough to calculate, often-times $P(\boldsymbol{x})$, which acts as a normalisation, is very challenging to determine. This is because calculating this normalisation involves integrating the likelihood over all the parameter space, and if there are a large number of parameters this can become enormously numerically complex. An alternative approach is needed!

That alternative comes in the form of *Markov Chain Monte Carlo*, MCMC. A Markov Chain is any mathematical system for which the next state of the system is dependent only on its current state; the system is in some sense "memoryless". For a large class of Markov Chains — those that are "ergodic" and "aperiodic" — one can prove that regardless of the initial position on the chain, the probability distribution converges to the same distribution []. MCMC uses such a Markov Chain which attempts to converge towards the posterior density distribution in particular. In MCMC, after choosing the initial position in the parameter space, successive states are chosen at random with a probability dependent only on the properties of the current position in parameter space and the proposed position. The convergence property of Markov Chains means that the set of steps made in the parameter space after some initial number of steps will have a distribution that converges to that of the posterior density distribution.

There are a number of MCMC algorithms, and the particular one used in this analysis is that of the *Random-Walk Metropolis-Hastings Algorithm*. In this algorithm, after the initial position in the parameter space $\boldsymbol{\mu}$, a new step is proposed some distance from the current one, $\boldsymbol{\mu}'$. This step is chosen at random from a multivariate Gaussian

distribution centred on the current position, with widths in each dimension of the parameter space chosen beforehand as constants for tuning the MCMC process. This choosing of a new proposed step at random is what gives the algorithm its Monte Carlo and Random Walk titles. Once a new step is proposed, it is accepted as the new step with a probability $S(\boldsymbol{\mu}'|\boldsymbol{\mu})$ according to the condition of *detailed balance*:

$$\begin{aligned} S(\boldsymbol{\mu}'|\boldsymbol{\mu}) &= \min\left(1, \frac{P(\boldsymbol{\mu}'|\boldsymbol{x})}{P(\boldsymbol{\mu}|\boldsymbol{x})}\frac{R(\boldsymbol{\mu}|\boldsymbol{\mu}')}{R(\boldsymbol{\mu}'|\boldsymbol{\mu})}\right) = \min\left(1, \frac{L(\boldsymbol{\mu}'|\boldsymbol{x})\,P(\boldsymbol{\mu}')}{L(\boldsymbol{\mu}|\boldsymbol{x})\,P(\boldsymbol{\mu})}\frac{R(\boldsymbol{\mu}|\boldsymbol{\mu}')}{R(\boldsymbol{\mu}'|\boldsymbol{\mu})}\right) \\ &= \min\left(1, \frac{R(\boldsymbol{\mu}|\boldsymbol{\mu}')}{R(\boldsymbol{\mu}'|\boldsymbol{\mu})}\exp\left[\mathcal{L}(\boldsymbol{\mu}'|\boldsymbol{x}) - \mathcal{L}(\boldsymbol{\mu}|\boldsymbol{x}) + \ln\frac{P(\boldsymbol{\mu}')}{P(\boldsymbol{\mu})}\right]\right). \end{aligned} \tag{5.6}$$

$R(\boldsymbol{\mu}'|\boldsymbol{\mu})$ is the probability density that position $\boldsymbol{\mu}'$ is proposed as a step from position $\boldsymbol{\mu}$, and vice versa for $R(\boldsymbol{\mu}|\boldsymbol{\mu}')$. In most cases, because of the use of the same multivariate Gaussian in choosing proposals, $\frac{R(\boldsymbol{\mu}|\boldsymbol{\mu}')}{R(\boldsymbol{\mu}'|\boldsymbol{\mu})} = 1$ simply. This component only becomes important at the edges of the parameter space, preventing the sampling probability being incorrectly impacted if a proposed step goes outside the allowed parameter space.

It is the detailed balance condition that ensures convergence of the MCMC algorithm to specifically the posterior density distribution. Crucially, because it is only dependent on the ratio of posterior densities, the hard-to-calculate normalisation $P(\boldsymbol{x})$ in both posterior terms cancels out, meaning one only needs to calculate the likelihood and priors for each step, as well as $\frac{R(\boldsymbol{\mu}|\boldsymbol{\mu}')}{R(\boldsymbol{\mu}'|\boldsymbol{\mu})}$.

- Want to maximise the sensitivity to measuring these two parameters. Energy spectrum of signal is background-free above $\sim 5\,\text{MeV}$, but rate is substantially larger at lower energies. If one can minimise backgrounds at the lower energies, then there should be hopes of obtaining a measurement with greater precision!

- Set up analysis approach: a Bayesian analysis using MCMC. Explain why this was chosen at a high level first: allows for us to perform a relatively complex analysis with multidimensional PDFs, numerous backgrounds, systematics, constraints on

parameters, all whilst allowing us to obtain well-defined measures of uncertainty on our final results.

- Test statistic: binned extended log-likelihood. Explain why this is fundamentally the "right" test statistic to use. Allows for handling of constraints and systematics.

- Give an overview of how an MCMC works. Key idea: exploring parameter space in such a way as to reproduce posterior density distribution. Clever! Helps to avoid "curse of dimensionality" that often arises in fits with numerous parameters. Must also explain how we work in a fundamentally Bayesian, not frequentist, statistical approach.

- Explain cuts that I am using for the analysis. Demonstrate how these attempt to maximise our signal sensitivity. Leads nicely into choice of 2D fit in both energy and radius (cubed).

- Describe implementation of neutrino oscillations within the fitting procedure: flat priors on the oscillation parameters, but a strong constraint from the solar flux. Explain choices of constraint that are possible. MCMC varies oscillation parameters and flux scaling factor, which then modifies the solar signal PDFs through a de-facto systematic that is a function of true neutrino energy, a 3rd "bookkeeping" dimension in the signal's PDFs only. Neutrino oscillations simulated via PSelmaa, which accounts for MSW effect in both the Sun and the Earth. For computational speed, at run-time we actually use a lookup table with linear interpolation for the survival probability as a function of parameters.

- Implementation of systematics: we handle them generally as linear transformations acting on the vector of bin data. Clever! Which systematics do we expect to be particularly important for this analysis? Well, mismodelling in detector optics etc. can lead to changes in the measured energy spectrum of processes,

which can be decomposed into an energy scale term and an energy smearing term to first order. Systematics also possible in the radial dimension (expected to be less important?)

5.2 Analysis on Scintillator-Phase data

- Description of dataset chosen for analysis: full-scint data that satisfies the "gold" list of run selection requirements, between 1st June 2022 and March 2023. Starting date chosen to ensure radon levels have stabilised within the centre of the detector.
- Impact of cuts on data and MC. Show tables (the full details maybe in an appendix) indicating this.
- Describe the constraints chosen to apply to the fit, and why they can be justified.
- Running & validation of MCMC fitting. Show plots of parameter values versus step, to demonstrate that the step sizes have been tuned sufficiently. Show auto-correlation plots, to motivate a sensible "burn-in" size. Show posterior density plots for each nuisance parameter, to check that they all look sensible and have sufficient statistics. Show plot of correlation coefficients between parameters, and look at any correlations that are particularly interesting.
- Look at the data versus MC plot in energy, radius, and both. Is there a good fit to data? Any clear disagreements?
- Show 2D contour plot for oscillation parameter posterior density. Note salient features. Show 1D posterior densities for each oscillation parameter. Derive measurement result for θ_{12}.
- Show impact of modifying certain constraints on the final results of the measurement of θ_{12}.

5.3 Sensitivity Projections

5.4 Conclusions

References

[1] E. Turner, *A Measurement of Scattering Characteristics of the Detection Medium in the SNO+ Detector*, DPhil Thesis, University of Oxford, 2022.

[2] K. Majumdar, *On the Measurement of Optical Scattering and Studies of Background Rejection in the SNO+ detector*, DPhil Thesis, University of Oxford, 2015.

[3] S. S. Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, The Annals of Mathematical Statistics **9**, 60 (1938), Publisher: Institute of Mathematical Statistics.