**Climate
of the Past
Discussions**

# *Interactive comment on* "Automated ice-core layer-counting with strong univariate signals" *by* J. J. Wheatley et al.

**J. J. Wheatley et al.**

j.wheatley@shef.ac.uk

**General comments:**

**Referee:**

*Annual layer detection and counting is a key tool for ice core research and similarly varve counts, tree ring counts etc. There is a need for automated method for annual layer counting and evaluation of the uncertainty of the counting procedure, and as such, the study is appropriate for publication in CP.*

*The manuscript presents a clever way to automate annual layer counting by splitting the signal in question into "easy" and "hard" sections. In the "easy" sections, annual layer*

C1254

*boundaries can be assigned in an unambiguous way without much difficulty, while the "hard" sections are highlighted for manual evaluation or handled by assigning probabilities based on annual layer thickness statistics. The automated assignment of annual layer boundaries in the "easy" section is convenient, but will probably in the case of long data series only represent a modest improvement and/or work saver, as it not a difficult or laborious task to assign annual layer makings to a single data series with a well-characterized and strong annual signal. The real value of the methods thus lies in its ability to perform well when applied to the "hard" sections, or "issues" in the nomenclature of the manuscript, and especially its performance in difficult case, e.g. when either the annual signal is unclear or when the data has problems with marginal resolution, frequent missing data sections etc. Unfortunately, the results presented do not allow a full assessment of whether the method performs well under these more difficult conditions.*

**Response:**

Firstly we would like to thank the referee for a comprehensive assessment of our work which raises a number of interesting points. We aim to address them below, in part through the provision of 3 additional analyses—we will include a summary of these in the revised paper.

In our manuscript we present a method of splitting a signal with a strong annually cyclic component into non-overlapping sections: 'certain' runs (or parts of a cycle) where the seasonality is clear cut and deterministic count can be made; and 'issues', sections where more attention is required. This classification process is very simple in terms of computation and is therefore very quick, taking less than a second on a modern lap-top for the Gomez core. It is consistent, repeatable, and does not require any prior manual assessment of chronology. Arguably the real strength of our approach as described in the manuscript is this robust, efficient separation of what the referee calls 'easy' and 'hard' sections.

The presented method relies on a sufficiently large sampling rate and annual cycle-to-noise ratio. In the case of an unclear (or weak) annual signal, or when resolution is low compared to the length of an annual cycle, it is the classification process that may require modification. Once issues have been identified they can either be presented for manual assessment (with the aid of simple reconstructions and probabilities based on their length) or a more sophisticated automatic process could be employed.

**Specific comments:**

**Referee:**

*The method performs well when applied to the test data presented, but that is not very surprising given the unusually fine data quality – the sampling rate is very high and the noise/non-annual part is weak — and the exceptional simple, regular and well expressed annual signal.*

*As I state in the general comments, I think the method is clever and has potential, and I think that it should be published, but I think the current manuscript is almost too pretty and too much based on low-hanging fruit to allow the reader to evaluate the potential strength of the method: Given the quality of the data and the limited length of the test data section (153 years), any reasonably successful method should be able to produce a count close to the target. Thus, the true potential of the method remains to be demonstrated. In short, it would be interesting to test the method's performance on data with less favourable sampling rate and on data with a less clear-cut annual signal.*

**Response:**

The example in our manuscript is a dating of the whole Gomez core—not a short well-behaved portion. We use the $H_2O_2$ signal as it clearly has the strongest annual

component and is therefore most reliable, and is therefore the most realistic choice within this data-set. We will clarify this is the revised paper. A further benefit of $H_2O_2$, whose cycles are directly correlated to sunlight, is that with an appropriate choice of $\nu$ the 'certain' runs are analogous to seasons. Our chronology therefore allows the analysis of other signals in seasonal terms. The shortness of our example section is not an advantage to the method (although of course it reduces the computational time)—a longer stretch of data would provide more information with which to model the 'certain' run lengths.

One challenge posed by the Gomez $H_2O_2$ signal, beyond the exponential reduction in layer thicknesses, is that the resolution at the top of the core is very high, we see regular 'fluctuations' in the annual cycles, whereas the bottom of the core has very few complete cycles due to the regular stretches of missing values.

The referee states that "any reasonably successful method should be able to produce a count close to the target". We are not aware of published successful methods in this sense, and we emphasize that the count from our method, as summarised in figure 8 of the manuscript, is far from being its only output. The method also provides a choice of layer boundaries, with uncertainty measures, from which change-points, peak apexes, and trough nadirs can be calculated. In the Gomez example troughs are placed in one-to-one accordance with, and in very close proximity to, those assessed manually.

**Referee:**

*With regard to the sampling rate, a data series with $\sim 8$ samples per average year would be a good test. If the authors have no other suitable data for a test along these lines, a down-sampled version of the data used in the study could be used.*

**Response:**

The presented classification process requires at least one point per 'certain' run; there are four runs per cycle which are not assumed to be of equal length. It is therefore the

shortest runs of the shortest cycles in the signal that need to be considered here. A data series with $\approx 8$ samples per average year is likely to have a lot of runs missing, causing a great number of issues that consist of the shortest cycles. This would be fine if issues were presented for manual assessment, however the modelling process would adversely affected as only the medium and long runs' lengths will be classified and modelled.

In the additional analyses, we do now explore the effect of sampling rate on data from the NGRIP core, where there is more variability in cycle length than seen in the Gomez $H_2O_2$. Here an average of $\approx 16$ points per cycle appears to be the limit for classification and probability assignment as presented in the manuscript.

**Referee:**

*With regard to the complexity of the annual signal, there are several possible tests that could substantiate the results: – Using data with a more complex annual signal, for example ECM data, water isotopes, or the (challenging) Visual Stratigraphy data of Winstrup 2011). Addressing the method's sensitivity to annual layer thickness variability, which seems unusually small in the data set used, potentially because the data are from a high accumulation site which has many annual precipitation events. Annual layer statistics could be used to evaluate if the test data used in the manuscript have unusually low layer thickness variability, i.e. when compared to data from records with thinner annual layers (both Rasmussen et al. 2006 and Andersen et al. 2008 already referenced presents relevant Greenland statistics), and a comparison to e.g. WAIS layer statistics would also be relevant if available.*

**Response:**

Our method requires a signal with a strong annual component; the classification process would require some modification for far noisier signals such as ECM or Visual Stratigraphy. For data with a high sampling rate, pre-smoothing the signal with a short interval moving average my be sufficient. We include $3$ additional analyses; the first $2$

apply the method exactly as presented in our manuscript to signals from the NGRIP ice-core (Greenland) which are a little noisier and have more varied layer thicknesses. The third example is a more challenging signal from the Gomez core.

In the manuscript, the probabilities assigned to the uncertain sections come from a regression analysis of the 'certain' runs and therefore not directly from whole layer thicknesses. Here we provide an analysis of 'trough to trough' cycle lengths (see figure 9 in the paper) to address the question around layer thickness variability in the Gomez core. This core covers the firnification process, and an exponential decay in layer thickness is observed. There are several cycles in the first $20$m of the core that contain over $100$ data points ($2$m), in contrast the last $5$m of the core has several cycles with fewer than $20$ data points ($40$cm) - a five fold decrease. This is well modelled by a linear trend on the log transformed thicknesses fit via simple linear regression, and the resulting model shows good homoscedasticity (constant variability):

$$\ln(\ell_i) \sim N(\ln(\hat{\ell}_i),\ 0.035)$$

where $\ell_i$ is the observed layer thickness at depth $i$. The mean layer thickness (and $99\%$ C.I.s) under this model for the start (@ $3$m), middle (@ $65$m), and end of the core (@ $132$m) are respectively: $171$cm ($109$cm, $268$cm), $94$cm ($59$cm, $147$cm), and $49$cm ($31$cm, $77$cm). Note the asymmetry in the confidence intervals. To allow comparison with the statistics in Rasmussen et al. (2006): under this model the probability of a random annual layer being either double or half the mean thickness at any given depth is $\approx 0.025\%$.

In the case of very high local variability in layer thickness this method may need some modification at the standardisation stage. One possibility would be to use a longer interval to estimate $\mu$ and $\sigma$, say $2$ or $3$ cycle lengths. Once the signal has been standardised satisfactorily, the method of classifying into 'certain' runs and issues is very robust - it is 'blind' to cycle lengths and looks only at the seasonal pattern in the signal.

The exact same method (code) identified the $> 2$m cycles and the $< 40$cm cycles. The high local variability in layer thickness would be modelled via the regression, causing higher uncertainty in the resulting chronology.

**Referee:**

*Another comment relates to the issue reconstruction probability assignment on and around p. 2487/14: The method as described and Fig. 5 indicate that some of the data that actually are available are disregarded as the probabilities are based solely on layer thickness statistics. Maybe the authors could think of a clever way to utilize the (dis)similarity of the different reconstructions and the available data across the "issue" (i.e. the black curve bits on Fig. 5) to refine the reconstruction probability assignment.*

**Response:**

The probabilities are calculated by modelling the 'certain' runs and applying the model to the issue lengths. However information about the classifications of the 'certain' runs that bound the issues and 'potential' runs within the issues is also used. This method is intended to be simple and quick and as a result does not make use of the data points within an issue.

Under the additional assumption that (a) time is equally spaced within the runs and (b) the standardised signal can be described by a sine curve, the sum of squared differences between the data and the reconstructions could be used. This is of little use in the case of missing values but may be useful to flag issues where the most probable reconstructions differ greatly from the standardised signal.

**Referee:**

*Also, the way issues is handled statistically in section 4 implies that the true curve shape (disregarding sampling problems and missing data) is a sine. I fear that this way to interpret sections of difficult data is too simplistic, especially in the case where the difficulties could also be cause by a less well-behaved signal. This calls for tests*

*like the one indicated above, and/or a discussion of how the method or pre-processing steps could/should be adapted to different data characteristics.'*

**Response:**

When assigning probabilities, the way that issues are handled depends only very weakly on the idea of an underlying sinusoidal curve. The key assumption is simply that the run classifications must follow the $P\ D\ T\ A$ pattern. In addition, in the log-linear model for run lengths, runs are fitted at depths implied by the sine reconstruction (equally spaced in the case $\nu = 1/\sqrt{2}$). This could readily be relaxed, e.g. by fitting all run lengths at the central depth of the issue, with little effect on the resulting probabilities, using the sine reconstructions only as visual aids for manual assessment. We will make this clearer in the revised paper. (See also the response to related minor issues.)

**Referee:**

*Finally, I miss a discussion of the potential (if any) to extend the method to multiparameter data.*

**Response:**

If the classification method can be extended to a multivariate framework—using multiple signals to group depths into 'certain' runs described by length and label—then the same method of assigning probabilities to issues can be utilised. In the univariate example we split the signal into three groups along the real line; it is possible split two out-of-phase standardised signals into four groups along the bivariate plane. Plotting $x'$ against $y'$ the points can be collected into quadrants: $Q_1$ where $x' > 0$ and $y' > 0$, $Q_2$ where $x' < 0$ and $y' > 0$, $Q_3$ where $x' < 0$ and $y' < 0$, and $Q_4$ where $x' > 0$ and $y' < 0$. These could be treated as potential runs in the manuscript—the method would be the same from that point.

**Referee:**

*The authors mention in the very last lines of the conclusion that work is under way to address some (if not all) of the above-mentioned issues, but I really miss some of these results and discussions in the current manuscript.*

**Response:**

The work mentioned uses a model-based statistical approach to make inferences about detailed chronology. It is much more computationally demanding than the method described in the manuscript, and is best seen as being complementary to the method in the manuscript, rather than an extension or alternative to it. One possible strategy would be to use the method of the manuscript to define and investigate issues; those issues that cannot be readily resolved (i.e. where no single reconstruction is obviously correct, based on run-lengths) could then be analysed in more detail, using the model-based approach to refine the chronology and the probabilities. We will provide an improved discussion of this point in the revised manuscript, however details of the model-based approach are beyond the scope of the present paper.

**Referee:**

*The manuscript text is not very long, but still describes the details of the rather simple method in at least sufficient detail. The abstract is fine and the language as well as the artwork and technical quality is good. Relevant annual layer detection / counting work is referenced, and the briefness of the reference list mainly reflects that automated annual layering methodology is an emerging field. Maybe the authors would like to elaborate a bit on how their method assumptions compare to those applied by some of the referenced works and to varve data as part of the BMPix tools of Weber et al.. The number of figures is on the high side, but many can be reproduced in small size if the legends etc. are sized appropriately. Fig. 4 carries little information in itself, as almost everything is reproduced in Fig. 5. Also, Fig. 6 is not essential. Figs. 10 and 11 could be integrated as the captions are almost identical.*

**Response:**

Thanks, we spent some time on developing a succinct description of our method.

The assumptions behind those alternative methods which we cite in the manuscript are rather difficult to formalise, generally being implicit in an algorithm and/or the human input to a semi-automated method. In the course of the responses to the referee's other comments, we have tried to clarify the assumptions that are or aren't made in our method; one of the additional analyses also relaxes some of these assumptions.

We are grateful to the referee for bringing the PEAK tool of Weber et al. (2010) to our attention. Their three methods are applied to high resolution signals induced from scanned images of tree rings, marine varves, and marine laminae, and have some points of similarity with ours. The 'zero-crossing method' algorithm iteratively finds the points at which the (raw) signal crosses a wide interval Gaussian moving average—with great effect even on noisy data. The 'frequency truncation method' algorithm similarly finds the zero-crossing points in the signal after high-frequency noise and low-frequency shifts have been removed via Fourier transformation. In the nomenclature of our manuscript, these methods essentially segment the signal into 'certain' runs representing peaks and troughs. They also present the 'maximum count method' algorithm which iteratively picks peaks as local maxima along a smoothed signal (short interval). Each of these methods have user defined parameters that represent minimum layer thickness and a 'minimum amplitude' tolerance which are adjusted to tune the count visually, along with the smoothing and frequency parameters. These methods provide a point estimate layer count, along with valuable information about the positioning of layer markings and the cycle lengths. Their most recent paper does not address missing values or provide a measure of uncertainty. We will include a discussion of the work of Weber et al. (2010) in the introduction of the revised paper.

**Referee:**

*In conclusion, I think that the current version presents a nice and clear description of 1. a few elegant and well-chosen data pre-processing steps, 2. a simple and efficient way to characterize the annual signal in the easy parts of the record, 3. a simple way to assign probabilities to different number of annual layers across difficult / reconstructed parts of the record based on layer thickness statistics 4. the results themselves and comparison to the results of the manual count. 5. sufficient test of parameter sensitivity etc. However, the manuscript fails to sufficiently demonstrate that the method represents a significant advance because the test data chosen are not sufficiently challenging. With very few news among the elements of the presented method (possibly with the exception of the assumption 2486/3) and no demonstrated performance on difficult data, I think the calorie count is on the low side.*

*To add some weight, I therefore suggest that - the method in its current form is applied to a more challenging data set in order to test whether the method represents a significant improvement compared to much simpler methods (e.g. by addressing some of the comments and suggestions above), or - the manuscript is seen as the first part of a manuscript that in its second part will elaborate on at least some of the material which is mentioned as ongoing work in the outlook In the first case, the manuscript's score on Scientific Significance will likely increase to 1 or 2 and would be recommendable for publication with minor technical/revisions.*

**Response:**

We include some additional analyses aimed at addressing some of the comments and suggestions above. We are not clear what the "much simpler methods" mentioned are, but if this refers to other methods that we have cited, we believe that, with these additional analyses, we have demonstrated the benefits of the approach in our manuscript.

**Technical Corrections / Minor issues:**

**Referee:**

*2478/22: While it is clear that noone has yet presented a statistically rigorous treatment of uncertainty of annual layer counting, it's not quite fair to say that "Little consideration" has been given to the issue.*

**Response:**

Yes, this sentence in the manuscript is misleading: the literature provides a good discussion of uncertainty. What was meant is that previous automated methods, Winstrup (2011) excluded, provide no measure of uncertainty on their chronologies. That is, we meant that the *models* show 'little consideration', not the authors. We will re-write this sentence, thanks.

**Referee:**

*2479/26: Why a sine wave? Even though the initial UV forcing may be close to sinusoidal, the log transform should change this, and unless the mean is thought to vary within each annual cycle (in which case any signal with a quasi-periodic behavior can be thought of as a sine wave on a non-linear time-scale with varying amplitude and mean), there is no basis for assuming that the signal resembles a sine. The statement needs clarification (e.g. what is meant by varying amplitude and mean) or can be removed.*

*2480/25: As above. No convincing arguments or evidence for the signal being sinusoidal is presented.*

*2486/18: This symmetry may not hold for other data sets, esp. after normalisation and log transformation. The authors could address what would happen in this case.*

**Response:**

The rationale for the sine wave is primarily empirical: after suitable (and simple) transformation, discussed below, the Gomez $H_2O_2$ signal does look approximately sinusoidal. This can be seen in figure 4 of the manuscript (and also in figure 2(c), though the compressed horizontal scale makes that more difficult). From the regression modelling, for all $\nu$ presented, $P$ and $T$ ('extreme') run lengths were found to be equivalent, $A$ and $D$ ('central') run lengths were found to be equivalent, and all run lengths have the same slope—this is not a requirement of the modelling but is evidence of a sinusoidal standardised signal. A further indication is that the lengths of runs of all classifications are found to be equivalent at $\nu = 1/\sqrt{2}$—the theoretical value at which this is expected of a sine wave. As mentioned in the response earlier, the precise shape of the underlying sine wave has little impact on the details of our analysis.

The standardisation and classification processes presented in the manuscript do very consciously assume a degree of symmetry in the seasonality. The raw Gomez $H_2O_2$ signal is asymmetric—its peaks are taller and spikier than its troughs. A logarithmic transformation is very natural here (not least because the raw data are non-negative) and induces symmetry with great effect; in other signals symmetry may be induced by different transformations, or in some cases transformation may not be needed. With very asymmetric data, one possibility would be to have two tuning parameters and classify potential peaks as runs of points above $\nu_1$ and potential troughs as runs of points below $\nu_2$. An alternative possibility is presented in the additional analyses, for Gomez non-sea-salt sulphur.

To clarify, we do allow the mean and amplitude of the cycle to vary within a cycle, but only slowly. In practice, this is implemented by estimating the mean ($\mu_i$) at each depth using a moving average over an interval length of approximately one cycle, and then similarly estimating the variability ($\sigma_i$) of the de-trended signal at each depth, again over approximately one cycle.

The referee is right that any signal can be thought of as a sine wave with varying amplitude and mean, *if* they were allowed to vary rapidly. Our use of slowly varying mean and amplitude allows us to preserve and extract the quasi-periodic behaviour in the data; the flexibility in this process is part of the reason that the method is not very sensitive to any underlying assumptions about the shape of the curve. The slowly varying amplitude is also the reason that, in the case of asymmetry, the relative heights of the peaks and troughs need to be taken into account, by one of the methods mentioned above.

We will clarify these points in the revised paper.

**Referee:**

*2484/3: . . . or sections where the annual signal simply isn't sufficiently clear cut.*

**Response:**

I think this refers to 2485/3. Agreed, thanks.

**Referee:**

*2485/20: Does this sentence end like it should?*

**Response:**

This could be replaced by: 'The *length* of a run, $\ell$ say, is . . .

**Referee:**

*2479/5: Winstrup now has a CPD reference that could be added.*

*2488/5: model's*

**Response:**

Agreed, thanks.

**Additional analyses**

Here we present $3$ additional analyses with the aim of addressing some of the points made above. The first $2$ example signals are from the NGRIP ice-core (Greenland) and apply the method, exactly as presented in our manuscript, to longer (and a little noisier) signals with more varied layer thicknesses. The effect of sampling rate per cycle on the count is also explored. The third example is a more challenging signal from the Gomez core, illustrating a possible way in which the run classification process can be adapted for asymmetric cycles. Additional figures A1–A12 are included in the supplementary material for illustration.

**NGRIP: ammonium and calcium**

Our method is applied to NGRIP Ammonium ($NH_4$) and Calcium (Ca) chemistry signals $(1440.49 - 1464.81 \text{m})$ which are sampled at 1mm intervals. We use the same algorithm as used on the Gomez $H_2O_2$ signal in the manuscript. These signals have a slightly greater noise to annual cycle ratio than the Gomez $H_2O_2$, with regular fluctuations and stretches of missing values.

Figures A1 and A2 show the effect of taking the natural logarithm of the $NH_4$ and Ca signals (respectively); in both cases (a) is the raw signal and (b) is the log signal. Note how the symmetry of the annual cycles is improved in both signals.

To test the effect of sampling rate, we run the analysis on three thinned down versions of both signals: taking every second point (2mm); every third point (3mm); and every fourth point (4mm). We refer to these as the second, third and fourth thinnings. An alternative way to generate signals of lower sampling rate would be to take averages of *non-overlapping* intervals (not a moving average)—we would expect our method to work better in this case.

There is very little trend in cycle length through this depth range in the NGRIP core. In each case we estimate the overall average cycle length from the ACF of the entire signal to use as the interval length when calculating $\mu$ and $\sigma$, effectively setting $\beta = 1$. Figure A3 shows ACFs for each signal: (a) no thinning (estimated average cycle length 62 points); (b) second thinning (32 points); (c) third (21 points); and (d) fourth (16 points). In each case $NH_4$ is shown as circles and Ca as stars.

Figures A4 and A5 show the standardisation process of the log $NH_4$ and log Ca signals respectively. These correspond to figure 2 in the manuscript: (a) is the log signal, (b) is the de-trended signal, and (c) is the standardised signal.

Figure A6 shows a stretch of the classified log $NH_4$ signal (fourth thinning) for $\nu = 0.49$ with 9 cycles: (a) is the log signal with $\mu$ and $\mu \pm \sigma$ shown as dotted lines; and (b) is the standardised signal. Points within an issue are black, points within peaks are coloured red, descending points are orange, troughs are blue, and ascending points are green. Here there are two issues; the first is caused by a single missing value and the second by a fluctuation in the data (a 'double peak'). Note that there are 3 ascending runs of length $1$ (the reason why we could not do a fifth thinning).

Figure A7 shows a stretch of the classified Ca signal (second thinning) for $\nu = 0.53$ with 7 cycles: analogous to figure 6 in the manuscript. There is one issue—a probability of 10% is assigned to there being two troughs in this section.

The probability distributions for the resulting chronologies are summarised in figures A8 ($NH_4$) and A9 (Ca) over a range of eight $\nu$ values, these are analogous to figure 10 in the manuscript: (a) is with no thinning; (b) is the second thinning; (c) third thinning; and (d) fourth thinning. The range of $\nu$ was chosen in each case as the interval over which the count is 'most stable'—in that the probability distributions are most similar. In each case a cursory check on the model's choice of 'certain' runs, and the probabilities assigned to resulting issues, was made to confirm that they are sensible. These ranges were chosen by eye and vary for each thinning; for the most part this is

due to the effect of the sampling rate on the visibility of the fluctuations and identification of annual cycles. Choices of $\nu$ below these intervals generally overestimate the count—classifying fluctuations as 'certain' cycles. Choices of $\nu$ above these intervals generally underestimate the count—missing out whole cycles in the classification of 'certain' runs. One way to stabilise the 'certain' cycle count would be to do one run of the classification process, model the 'certain' run lengths, check for outliers in the distribution (abnormally short or long runs), and assign these as issues. Note that the Ca count is generally higher than the $NH_4$ count and has a greater uncertainty, this is due to a number of extra potential annual cycles present in the Ca signal when compared to the $NH_4$ signal, suggesting that a bivariate implementation, as outlined above, would be beneficial.

For comparison, after correcting for the slight decreasing trend in the 'trough to trough' log cycle lengths from the most probable $NH_4$ chronology and modelling them as Gaussian, the probability of a random annual layer being either double or half the mean thickness at any given depth is $\approx 2.3\%$.

**Gomez: non-sea-salt sulphur**

Alongside the other challenges of the Gomez core, its non-sea-salt sulphur (nss-S) signal has an additional complication. The shapes of the annual cycles change with depth—from wide noisy troughs at the top of the core to cycles similar to those of the the $H_2O_2$ / $NH_4$ / Ca signals at the bottom of the core. A logarithmic transformation improves symmetry at the bottom of the core but not at the top. Figure A10 shows the nss-S signal for the entire Gomez core, which is not transformed for this analysis.

The standardisation method presented in the manuscript effectively estimates a local mean and standard deviation for each depth—points which exceed a given number ($\nu\sqrt{2}$) of standard deviations from the mean are classified as *potential* peaks and

troughs. As symmetry can not be induced throughout the nss-S signal we require a more robust method of classification—we use local quantiles or percentiles. Along interval lengths estimated from the ACF of the nss-S, as before, we calculate the local $v_1$th and $v_2$th quantiles at each depth, data points above the $v_1$th quantile are classified as potential peaks and data points below the $v_2$th quantile are classified as potential troughs. From this point on the method continues as in the manuscript, using $\beta = 10$ throughout.

In the regression model for the nss-S log 'certain' run lengths the $A$ and $D$ ('central') classifications are found to be equivalent, whereas $P$ and $T$ classifications show a statistically significant difference ($p < 1\%$). There is also a statistically significant interaction between the depth index and the classification factor ($p < 1\%$). This fits a steeper gradient to the trough log lengths relative to the other classifications—modelling the changing cycle shape down the core.

Figure A11 shows sections of classified nss-S with 7 cycles: (a) at the start of the core, and (b) towards the end of the core.

Figure A12 shows the resulting probability distribution for the chronology at $v_1 = 0.85$ and $v_2 = 0.5$, the most likely chronology ($p = 0.4$) has a one-to-one trough correspondence with the most likely reconstruction found from the $H_2O_2$ signal. This is true over the range $0.8 \le v_1 \le 0.9$ and $0.4 \le v_2 \le (v_1 - 0.3)$.

**Concluding remarks**

The manuscript attempts to show that this method is robust to the tuning parameters. Plots are provided to show the effect of changing $\nu$ on the resulting probability distributions for the cycle count. In practice, since this method does not aim to give definitive probabilities, it could be argued that the robustness of the classification method, and of the issues found, is of more importance. In all four analyses, different issues arise

when varying $\nu$ in the ranges presented; see manuscript figure 3 for an example. However, in all cases where one value of $\nu$ has an issue over a depth range where another value of $\nu$ gives a 'certain' count, the reconstruction which corresponds to the 'certain' runs always has very high probability ($> 0.95$) and in most cases is assigned a probability of $1$ after normalisation. Thus the key message about which parts can be confidently classified, and which are genuinely uncertain, is highly robust. A discussion of this point will be included in the revised paper.