Referee comments are shown in italics below, and our responses are indicated by ">".

**Reply to referee #1 (E. Wolff):**

*Some of my comments are not really criticisms of the paper, but are more in the style of philosophical musings on the method, that might provoke some interesting discussion, or an added sentence.*

> Thank you for the very positive feedback. You are raising many valid points concerning several aspects of the method. Our responses are given below.

*Introduction: I think more should be made of the importance of getting an objective error estimate. For me, this is the crux of the improvement automated methods allow – if I make a range of reasonable decisions about what constitutes a layer, what is the range of ages I will reach? This is what manual counting simply cannot do, because it would require each sequence to be counted many times. Could you add more about that?*

>We have added the following: As the approach is based on statistics, it also allows for objectively inferring an uncertainty estimate of the resulting timescale. Such uncertainties are generally very hard to assess manually.

*Page 2521, line 15. I think you should clarify that while the method allows for multi-parameter counting, it has not yet been attempted.*

>This has now been clarified by adding the following sentence (p. 2521, line 16): … (although this is yet to be implemented).

*Page 2521, line 16. While we know from modern measurements that dust input to Greenland is seasonal (with one season showing highest values), we would have to admit that we cannot be absolutely certain this was the case in the last glacial. This assumption (that the same applies to the glacial) should be stated (and is an argument for multi-parameter).*

>We now state this assumption more clearly by adding: From modern-day ice core data it is known that the dust input to Greenland has a significant seasonal variation (Alley et al., 1997;Hamilton and Langway, 1968). Assuming this relationship to hold back in time, also the grey-tone intensity of the line-scan images should display a seasonal pattern.

*Page 2522, line 15-18. I don't think this paper requires such a long list of HMM applications.*

>Some of these have been removed.

*Page 2523, line 23: t should be inside the bracket, "(e.g. time, t) is considered"*

>The "t" is now placed before the bracket. We choose to place it here, as it does not necessarily imply time.

*Page 2525, line 11-13. This is crucial, because you are asserting that this is the basis of manual counting and therefore the basis for the automated method. I think it is wellstated here, and should be made quite prominent.*

>We have made this argument more prominent by ending the paragraph as follows: In this way the layer detection algorithm employs the same principles as used in manual layer identification.

*What is not so clear is how the two issues are played against each other, which may be the basis of subjectivity – how thick will I allow a layer to be before I decide a layer must be present even though the shape and amplitude are wrong? Is it worth saying something about this?*

>We have now mentioned the point that the two indeed are played against each other, and that their weight can be adjusted. However, in the current version, the weighting of the two has not been adjusted, which gives relatively larger weight to the layer signal than the layer length. The following sentence has been added to the manuscript (p. 2525, line 16): The relative weighting of these two criteria can be adjusted as desired.

*Page 2532, line 20 "deduced" not deducted.*

>This has been corrected.

*Page 2536. This whole idea of the sensitivity tests on artificial data is a good one, but is not described in enough detail here for the reader to judge. This section reads like a précis not a paper. I think you need to show a figure or else give a longer description. I don't think the reader would understand what you have done here. Page 2537, lines 1-10 is also hard to follow. Please re-write this section.*

>Two new figures have been added to illustrate the results of the sensitivity analysis, and the entire section has been revised to facilitate the reading. The figures illustrate the following:

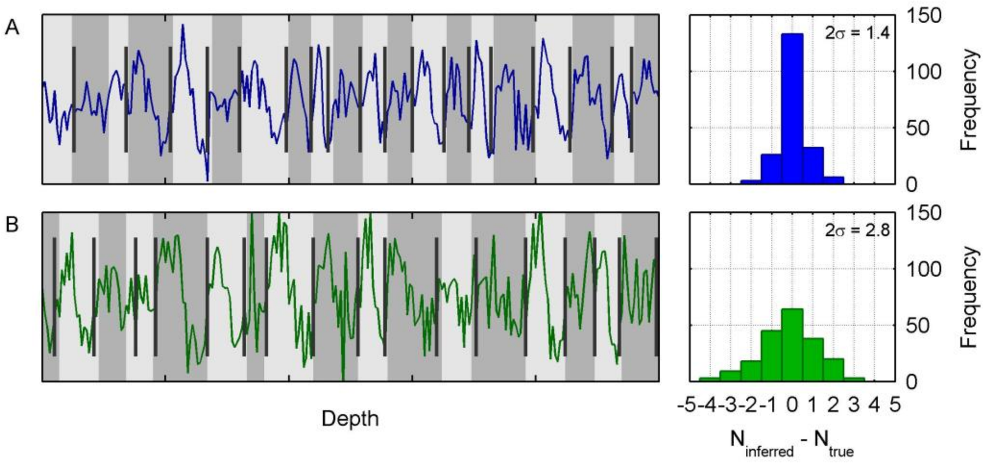

**Fig. 4:** Test of layer detection algorithm on ensembles of two types of synthetic data. **Left:** Example data series. Grey banding indicates the original layer boundaries, based on which the data has been produced. Dark grey bars signify the location of inferred layer boundaries. **Right:** Differences between inferred ($N_{inferred}$) and true ($N_{true}$) number of layers in the data series based on an ensemble of 200 data series with approximately 50 layers each. **(A):** Data is produced as a sequence of sinusodials generated by the parameters $\boldsymbol{\varphi} = 1$, $\boldsymbol{\Phi} = 0.5^2$, and $\sigma_\varepsilon = 0.5$ (Eq. 2). **(B):** As A, but with higher amounts of noise and variability in the layer signal ($\boldsymbol{\varphi} = 1$, $\boldsymbol{\Phi} = 1$, $\sigma_\varepsilon = 1$).

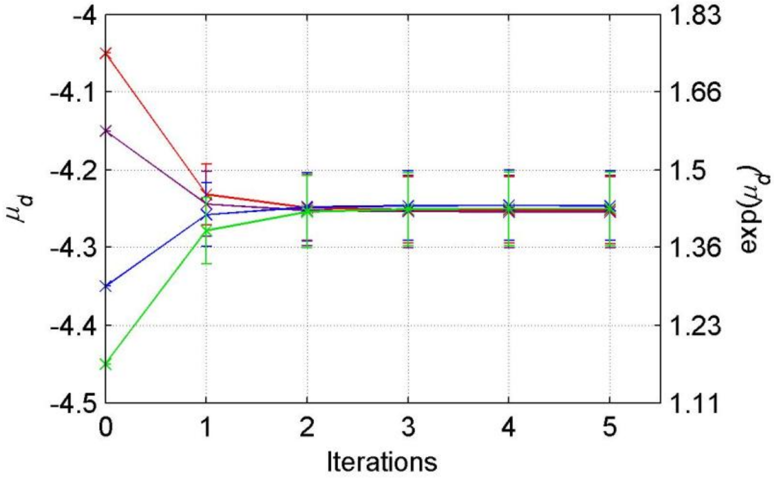

**Fig. 5:** Evolution of the derived layer thickness parameter $\mu_d$ as function of the number of EM iterations, when given incorrect input values of this parameter. Median of the resulting layer thickness distribution is $\exp(\mu_d)$. Data points are based on ensembles of 200 synthetic data series (sinusoidals produced with parameters $\boldsymbol{\varphi} = 1$, $\boldsymbol{\Phi} = 0.5^2$, and $\sigma_\varepsilon^2 = 0.5^2$), that each contains approximately 50 layers. Error bars signify the resulting 1σ spread within ensemble members. For the considered range of initial values, the derived value of $\mu_d$ converged to the original input value of $\mu_d$=-4.25 after just a few iterations.

*Page 2539. I don't think you should see it as the right test that you agree with the manual method. Rather I think this test shows that you have understood the manual method well enough to codify its assumptions. That means that now you are in a position to test how robust and important those assumptions are. With apologies to some of your co-authors, manual counters could be wrong! This thought actually leads on to the top of page 2540 when the optimistic uncertainty arises because you have not yet fully explored the range of parameters that might reasonable be altered.*

>I agree that a comparison with manual counting is not optimal, especially for a data section where the annual layering is ambiguous. However, for the line-scan data presented here, there exist no absolute-dated marker horizons, and only manual layer counts are available to test the method.

As also requested by reviewer #2 and the editor, we have now added to the paper a section where the layer detection algorithm has been applied to a data series ($\delta^{18}$O from Dye 3) with a more distinct annual signal. This data series was selected as it has been annually layer counted with high confidence, while also containing several absolute age markers in form of volcanic horizons. In this way, the performance of the method is also tested against this independent data set. The performance of the algorithm on this data set is within 1.4% of the manual layer counts, and agrees well with the time intervals between the volcanic marker horizons.

Furthermore, several paragraphs in the sections "introduction" and "results" have been added and rearranged in order to fully integrate the results from the Dye-3 data into the manuscript.

The results when applied to the Dye-3 data set:

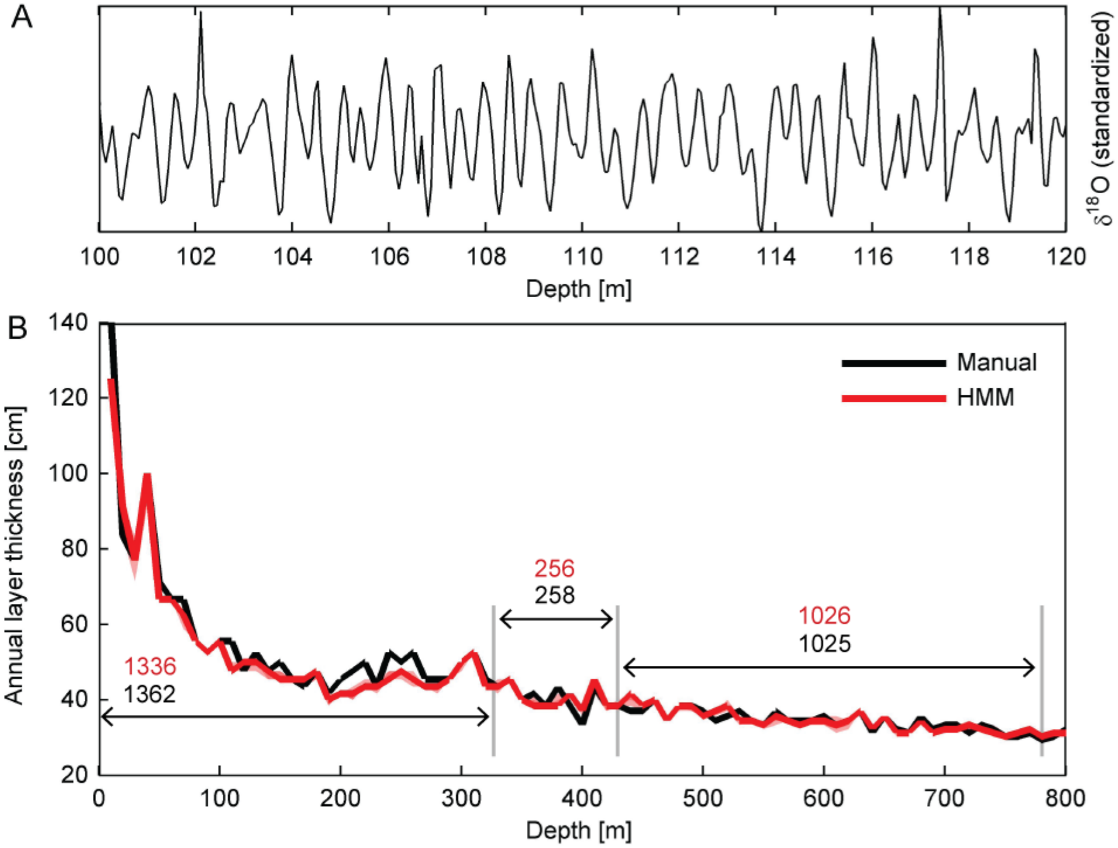

**Fig. 6: (A)** Example of the DYE-3 $\delta^{18}$O-record. The data was measured with approximately 8 samples per year. **(B)** Evolution in mean layer thicknesses for the upper 800 meters of the record, spanning approximately the last 2000 years (1979 - 79 CE). Annual layer thicknesses are calculated over 10 m sections. The 95% confidence interval derived for the automated counting is shown as a (very narrow) red shaded band. Uncertainties for the manual counting are negligible. The depth of three historically dated volcanic reference horizons (Öraefajökull, Hekla, Vesuvius) are indicated with grey bars. Red numbers denote automated layer counts within each section, which are to be compared to the historically known time durations (given in black).

**Reply to referee #2 (anonymous):**

We thank the referee for the review. Our answers are provided below.

*One of the constraints on the described approach is the assumption of stationary parameters within a model run. However, this assumption is addressed in part by implementing the algorithm as a set of sequential runs on short stretches of data. A wavelet or other frequency-based analysis of the dataset may be helpful for determining an appropriate segmentation of the dataset prior to dating with this method.*

> This is a good suggestion, and we have tried it out. However, a spectral analysis of the line-scan data does not yield any prominent peak corresponding to a 1 year oscillation. Also, it should be noted that an estimate of the mean layer thickness (in the form of $\mu_d$) is only used as a starting point for subsequent iterations. Based on the appearance of the layering in the data, the algorithm iterates using the EM algorithm, and is hereby able to find the most likely mean layer thickness by itself. An input in form of prior estimates of the mean layer thickness is therefore not necessary.

*If the PCA was calculated using the entire width of the image, as seen in Figure 1, then it is possible that selecting a narrower band of the image may provide better results. Variability in layer position across the scan that appears as wiggles in fig. 1 (or inclined layers elsewhere) would result in aliasing of these features in PCA axes 2 and higher. In some cases this might act as a type of edge detector that helps to define a layer, but that may not always be true. It would be interesting to know if a relatively narrow strip of values averaged by depth with less aliasing (but wide enough to average out bubbles, scratches, etc.) would perform as well as a PCA calculated from the full image width.*

>The intensity profiles were calculated based on a narrow band of the line-scan data to avoid such aliasing. The width of the band (50 pixels) was chosen such that it was reasonably wide to eliminate noise, but reasonably thin to avoid aliasing. However, given the very straight and horizontal layering in the considered depth interval, the dependency of the resulting grey-tone intensity profile on the exact width of the band was negligible.

*It is unfortunate that the analysis did not also include a shallow dataset with known volcanic tiepoints so that the accuracy could be assessed in more quantitative terms.*

>This has now been included (see above).

*I believe that some of the statements regarding the correspondence between the HMM method and the manual GICC05 chronology are overly optimistic (pages 2539-2540). The authors take an overlap in uncertainty bounds between the automated and manual counts as indicating correspondence. For example, it is stated that there are only two regions where the counts are outside the confidence intervals in figure 5c. Given the lack of statistical significance tests, it would be prudent to point out that the estimated thicknesses from each method are outside the others confidence interval more often than not. In the later discussion of GI-12 and table 1, the very slight overlap of tails between the automated and manual counts is mentioned as suggesting that the two are not entirely dissimilar. Given that the*

*GICC05 chronology is not independent of the NGRIP dataset, a more stringent interpretation that is more in line with a pair-wise test of variance might be appropriate. Table 1 suggests there would be an extremely high degree of significance in the difference. This does not definitely imply that one result is more correct than the other, but especially in the case of figure 5c, the fact that often just the extreme tails of the distributions are overlapping is important and needs to be acknowledged.*

>As the reviewer correctly points out, the two layer counts do not always line up. However, given the ambiguity involved in the layer counting for this data series, this is not surprising. Indeed, manual layer counting is not always consistent, especially when the layering is ambiguous. Thus, for the line-scan data used in the paper, a complete agreement between automated and manual layer counts is not to be expected.

We have changed some statements reg. the degree of similarity between the two, and to acknowledge the referee's point that in case of the counting in GI-12 only the tails of the distributions are overlapping, we have added the following sentence (p. 2540, line 20): [… the two associated 95% confidence intervals are overlapping], although only with their respective tails.

Generally, the uncertainty bounds inferred by the automated method are very optimistic, as it assumes unbiased counting and we have now emphasized this more clearly in the manuscript. In a section about the application to the Dye 3 data set, we have e.g. added the following:

Note, however, that the inferred uncertainty of the automated layer estimate is very small (0.4%) and does not include the true age. There are several reasons for this. One reason is that the algorithm has not been allowed to fully include the uncertainty that lies within the range of parameters that might reasonable describe an annual layer signal in the data series: The algorithm is allowed to select the most likely set of model parameters, but for the derivation of the corresponding confidence intervals, the uncertainty associated with this choice is assumed negligible. Additionally, the layer detection algorithm assumes the counting to be unbiased.

*This method shows very good promise for future development, and the authors have indicated useful directions to explore. Among these, a method of passing information from adjacent segments of the piecewise processing seems a high priority given their interpretation of the excursions in figure 5c.*

>Thanks. There is no doubt that a method of passing information from adjacent data segments is high priority. The implementation of this will not be difficult, but adds an additional number of parameters in order to quantify the degree to which adjacent data series should impact each other. For this reason, we have avoided doing so here.

*While comparisons between different studies are difficult, I note that in the conclusion the authors describe this approach as having "high skill," while their discussion in the background section was entirely dismissive of prior studies that had equal or better statistical results. It would be useful to put the current results more in context to prior work.*

>As previously mentioned, a section has now been added which shows the performance of the algorithm on a data series with a less ambiguous annual signal, where the performance of the algorithm is better.

It is, in our opinion, impossible to reliably compare the statistics of different studies using different data sets. The line-scan data is a difficult data series, both for manual and automated counting, and the disparity between these two are consequently relatively larger than for a data set of less ambiguous layering. We have mentioned several prior studies of automated methods for annual layer identification in the introduction, but have not included the statistics of these, as it in our opinion would be misleading. However, I think that a future study in which the performance of different algorithms on the same data series is investigated would indeed be very valuable.