A Hidden Markov Model Approach to Infer Timescales for High-Resolution Climate Archives

Winstrup, Mai

Published in:
Proceedings of the 30th AAAI Conference on Artificial Intelligence and the 28th Innovative Applications of Artificial Intelligence Conference, February 12 – 17, 2016, Phoenix, Arizona USA

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Winstrup, M. (2016). A Hidden Markov Model Approach to Infer Timescales for High-Resolution Climate Archives. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence and the 28th Innovative Applications of Artificial Intelligence Conference, February 12 – 17, 2016, Phoenix, Arizona USA* (Vol. 16, pp. 4053-4060). Arizone, USA.

A Hidden Markov Model Approach to Infer Timescales for High-Resolution Climate Archives

Mai Winstrup
University of Washington, 4000 15th Ave NE, Seattle, WA 98195, USA
Now at: University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen, Denmark
mai@gfy.ku.dk

Abstract

We present a Hidden Markov Model-based algorithm for constructing timescales for paleoclimate records by annual layer counting. This objective, statistics-based approach has a number of major advantages over the current manual approach, beginning with speed. Manual layer counting of a single core (up to 3km in length) can require multiple person-years of time; the *StratiCounter* algorithm can count up to 100 layers/min, corresponding to a full-length timescale constructed in a few days. Moreover, the algorithm gives rigorous uncertainty estimates for the resulting timescale, which are far smaller than those produced manually. We demonstrate the utility of *StratiCounter* by applying it to ice-core data from two cores from Greenland and Antarctica. Performance of the algorithm is comparable to a manual approach. When using all available data, false-discovery rates and miss rates are 1-1.2% and 1.2-1.6%, respectively, for the two cores. For one core, even better agreement is found when using only the chemistry series primarily employed by human experts in the manual approach.

1. Introduction

Over the last 2 million years, Earth's climate has oscillated between ice ages and warm periods with climate similar to present. The ice ages lasted approximately 100,000 years, while average duration of the warm periods was only 10,000 years. During the last ice age, immense ice sheets existed in the high northern latitudes, and covered large parts of North America. Approximately 11,700 years ago, the climate abruptly changed to current-day conditions, with the full transition completed within a few decades (Steffensen et al. 2007). By studying records of past climate, and investigating how and why climate has varied in the past, one can obtain a wealth of information on the intricate workings of the climate system. Such understanding is necessary also for making accurate predictions of future changes in a warming climate.

Timescales are fundamental to the utility of paleoclimatic records; without knowledge of the corresponding age, a measurement has little or no scientific value. Accurate timescales are needed for investigating periodicity and rapidity of past climate events, as well as for comparing climate records from different locations. They hold clues on the relative timing and spatial pattern of past climatic changes, which provides invaluable information on causes and mechanisms for these events. Consequently, major efforts in past climate research go into developing timescales for climate archives. The best method for constructing a timescale depends on properties of the particular archive, with depth and time resolution being key parameters. Under beneficial conditions, several types of archives (ice cores, sediment cores, tree rings, corals, etc.) may have sufficient resolution for annual layers to be identified, allowing a high-precision timescale to be constructed. We here focus on ice core data, but the method described is relevant also for other types of paleoclimate records.

The ~3km long ice cores drilled through the Greenland and Antarctic ice sheets contain excellent climate records informing about past temperature, atmospheric composition (including greenhouse gasses), volcanic and solar activity, among others. Each year, a layer of snow is deposited on the ice sheet surface, and it is gradually buried by the continuous snowfall. The snow quickly turns into incompressible ice, and gravity causes the ice to slowly move towards the ice sheet margins, where it eventually is removed by surface melt and iceberg calving (Fig. 1A). This process causes the ice layers to stretch and become thinner with depth, thereby progressively decreasing the temporal resolution of the ice core data. Greenland deep ice cores usually cover the last 100-200,000 years (see e.g. NEEM community members 2013),

while the longest Antarctic ice cores span more than 800,000 years but with correspondingly lower temporal resolution (Jouzel et al. 2007). Slowly melting sticks of the ice core, numerous high-resolution chemistry series are obtained by continuously measuring impurity concentrations in the melt-water stream, resulting in an effective data resolution of less than 1cm. If resolution permits, these ice-core chemistry series display annual cycles reflecting seasonal changes in atmospheric composition and circulation patterns. Black carbon, for example, is produced by forest fires and concentrations peak in summer, while sea-salt aerosols (e.g. Na) reflect the amount of surrounding sea ice and reach a maximum in winter (Fig. 1B).

Current practice in the paleoclimate community is to construct annual-layer timescales by laborious manual layer counting (Rasmussen et al. 2006; Andersen et al. 2006; Sigl et al. 2015a). The task is non-trivial; inter-annual variability in the ice-core chemistry series is large, and oftentimes the annual signal is obscured by external events. Volcanic eruptions, for example, give rise to highly elevated sulfur levels that overprint the annual sulfur cycles (Fig. 1B, lowermost data series). Preferably, layer decisions should therefore be based on parallel analysis of multiple chemistry series, while employing domain knowledge on seasonal timing of peaks and troughs in the various series and the variability of layer thicknesses. During development of the Greenland Ice Core Chronology 2005 (GICC05), a total of more than 60,000 layers were counted by hand (Svensson et al. 2008). Development of this timescale was a several-year effort for multiple experienced researchers; each investigator counted and recounted layers independently, and subsequently the layer counts were reconciled in discussion between all investigators. In total, an estimated 5-10 man-years were spent counting and reconciling layers (A. M. Svensson, pers. comm.). Inter-annotator agreement was sometimes quite poor, with discrepancies up to 10% over particularly difficult sections (Rasmussen et al. 2006). Uncertainty on the timescale was estimated by introducing so-called uncertain layers, each counted as ½±½ year. Uncertain layers were assigned when e.g. the annual signal was present only in some chemistry series, the relative timing of peaks in the various series was atypical, in sections with data gaps, or if agreement between investigators could not be reached (Andersen et al. 2006). The GICC05 uncertainties range from ±1-2 years over the last 2000 years, 1-2% during the last 10,000 years, and increasing to 5% thereafter.

There have been some previous attempts to automate the layer counting process (e.g. Wheatley et al. 2014; McGwire et al. 2011), but they have had limited success. A primary issue is the large irregularity of the annual layer signal, which confounds both manual and automated

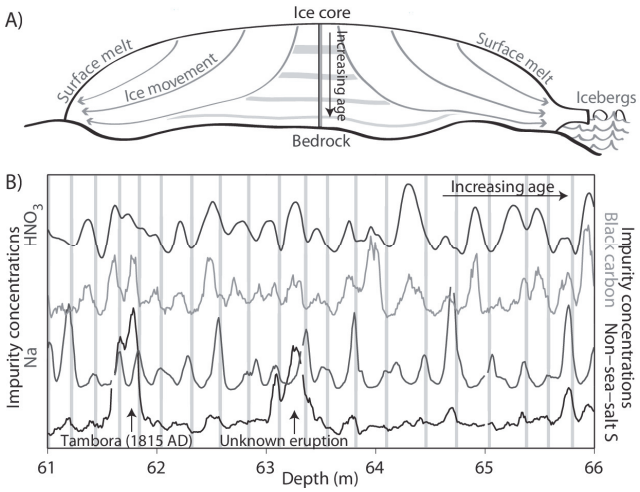

Figure 1: A) *Schematic ice sheet with an ice core. As ice slowly moves from the central parts of the ice sheet towards the margin, layers of ice (grey horizontal lines) are progressively stretched. An ice core therefore has thicker annual layers for recent times (near surface), and thinner layers with increasing depth and age.* ***B)*** *Chemistry series from a 5m section of the Greenland NEEMS1 ice core, covering the age interval 1795-1815 AD. StratiCounter layer counts (grey bars) mark the beginning of a calendar year. Large volcanic eruptions give rise to high sulfur concentrations (non-sea-salt S; lowermost data series). During this period, two large eruptions took place, including Tambora (Indonesia, 1815AD), traces of which are visible at a depth of 61.8m.*

approaches. In face of this, obtaining objective results that match the accuracy of manual layer counting by trained personnel is a real challenge. Most methods developed so far fall short of this goal, primarily because their framework is not designed to incorporate the wealth of explicit and implicit knowledge employed by human experts when counting layers in a core.

In this paper, we describe and demonstrate a Hidden Markov Model-based algorithm called *StratiCounter* that automates the arduous decision-making process of annual-layer counting. *StratiCounter* is computationally efficient; on a single chemistry series, it can count up to 100 layers per minute on a 2015 laptop, with further efficiency gain if expanded to parallel processing. The time spent increases linearly with the number of chemistry series. Ignoring various complexities (GICC05 was, for instance, constructed by piecing together sections from several cores), this would amount to the entire GICC05 timescale being produced in a few days, instead of a several year-long project occupying multiple researchers.

As input, *StratiCounter* uses multiple chemistry series with annual cycles and an initial set of layer boundaries; it outputs probabilistic age estimates along the core, together with the most likely layer boundaries. *StratiCounter* mimics a manual approach by incorporating domain

knowledge and mirroring manual layer-counting procedures. Main features are:

1) Appearance of annual layers in the data is described in terms of their varying shape and layer thicknesses.
2) It leverages information from multiple co-registered chemistry series containing an annual signal, thereby greatly improving the accuracy of the resulting timescale.
3) Layers are inferred simultaneously based on information contained in a complete core section, hence all data assist the correct identification of fuzzy layer boundaries.

To evaluate the performance of *StratiCounter*, we use manual layer counts as chronological reference, while recognizing that these layer counts are not really perfect.

A statistical layer-counting approach has several advantages over a manual approach. Obvious advantages are the lessened burden of manual labor and the objectivity of the outcome. An additional benefit is a rigorous definition of uncertainties. Few manually-counted timescales have uncertainty estimates, and those that do exist are very subjective. To compensate for the subjectivity, total uncertainty is often approximated as linearly increasing with depth. For example, the age uncertainty of GICC05 was derived by summing the contribution from all encountered uncertain layers. This was later found to be a very cautious uncertainty estimate that, while probably reasonable on shorter timescales (100-1000 years), grossly overestimates the accumulated age uncertainty at large depths (Buizert et al. 2015). Assuming the inherent uncertainty of the counting procedure to be a random process, the increase in uncertainty ought to resemble a random walk, i.e. increase as the square root of age. Uncertainty estimates from an objective procedure that is meticulously based in statistics have potential to reflect the accumulated age uncertainties at larger depths more accurately, and thereby reducing them. As a result, confidence intervals produced by *StratiCounter* are significantly smaller than their manual counterparts.

Given the subjectivity of manual layer counting, there is a risk of unconscious bias towards previously established timescales, which may or may not be accurate (Sigl et al. 2015b). Of course, bias may also occur in an automated approach, but the risk of unconscious bias is less. The researcher running the code is forced to specify assumptions explicitly, thereby encouraging scrutiny of their validity. If, for example, one is using exogenous arguments to constrain the algorithm to produce a specific age at a given depth, these assumptions should be consistent with an original unconstrained timescale to within the known uncertainties.

Algorithms excel in exploiting very large quantities of data, thereby allowing a more data-intensive approach than a human expert. The first manually-counted ice-core chronologies relied on the annual signal in a single data series (Hammer et al. 1978). With improvement of data extraction techniques, the importance of forming a coherent picture of the annual layering based on multiple chemistry series has become increasingly apparent (Rasmussen, Svensson, and Winstrup 2014). A recent revolution in measurement techniques has resulted in large quantities of high-resolution chemistry data available for layer counting. In one lab, for example, ice cores are routinely analyzed for 30+ chemical components (J. McConnell, pers. comm.). Most of these display an annual signal, and thus contain age information, but it is intractable to use them all during manual layer counting, so generally only the best 3-5 chemistry series are employed. Even a few series can be difficult to manually analyze in parallel, and consequently there is a tendency to select (not always consciously) a single chemistry series as the "master" and consult the remaining data only when in doubt. In contrast, an automated approach is able to extract information from the complete data, thus potentially improving the resulting timescale.

A single-chemistry-series version of *StratiCounter* was described previously in the geoscience literature (Winstrup et al. 2012). Here, we describe a mature version of the algorithm that can employ multiple chemistry series with annual signal.

2. The *StratiCounter* Algorithm

We use the term "annual layer" for the ice-core segment corresponding to a year, i.e. each data point is part of exactly one layer. This is in contrast to the general usage of the term in a manual layer-counting context, where it refers to a layer boundary. Rather than looking for discrete layer boundaries, we wish to infer the most likely layer at each depth along the core. This can be framed as a hidden-state problem, with the states being enumerated layers assigned to discrete depths, and it can be solved using the existing framework of Hidden Markov Models. This is the approach taken in *StratiCounter*.

At all depths, we wish to find the probability distribution of layer state, with each annual layer identified by consecutive numbering from beginning of the data series. Such probability distribution contains information on the most likely age, as well as a confidence interval for the age estimate. We use the notation $S_t = j$ to indicate that the layer state is j at discretized depth t, and denote by $\tilde{\gamma}_t(j)$ the probability of being in layer j at t, when conditioned on the observed data ($\boldsymbol{o}_{1:T}$) and employed model (θ):

$$\tilde{\gamma}_t(j) \equiv P(S_t = j | \boldsymbol{o}_{1:T}, \theta) \qquad \text{(Eq. 1)}$$

Layer thicknesses tend to be quite regular. Given the frequent ambiguity of the annual layer signals, this knowledge is important to include in the algorithm. We can impose a layer thickness probability distribution $p(d)$ as prior for the thickness of each layer, and efficiently calculate the probabilities in Eq. 1 using an adapted version of the Forward-Backward (FB) algorithm for Hidden Semi-Markov Models (HSMM). The specific variant of HSMM used here is also called an explicit duration HMM. This framework is useful since the problem can be reformulated as inferring the hidden state (i.e. layer number) of each data point, with the succession of states being a (fully predictable) Markov chain process, and with a prescribed prior for the duration of each state.

Two main inputs to the FB algorithm are the layer thickness probability distribution, $p(d)$, and estimates of the likelihood that data segment $\boldsymbol{o}_{t_1:t_2}$ exactly constitutes a complete annual layer. We use similar notation as Yu (2010); we denote this likelihood $b_j(\boldsymbol{o}_{t_1:t_2})$, and use two square brackets to indicate that the layer starts and ends exactly at t_1 and t_2, respectively:

$$b_j(\boldsymbol{o}_{t_1:t_2}) \equiv P(\boldsymbol{o}_{t_1:t_2} \mid S_{[t_1:t_2]} = j, \theta) \qquad \text{(Eq. 2)}$$

All layers are assumed to be produced by the same process, so the dependence on j can be omitted. The data point $\boldsymbol{o}_t$ may be a vector, and may thus contain values corresponding to multiple chemical species measured in the ice core at depth t. In section 2.2, we describe a way to compute these probabilities.

The general forward (α) and backward (β) equations for explicit duration HMMs are (Yu 2010):

$$\begin{aligned}\alpha_t(j,d) &\equiv P(S_{[t-d+1:t]} = j, \boldsymbol{o}_{1:t}|\theta) \\ &= \textstyle\sum_{i\in\mathcal{S}\setminus\{j\}} \sum_{d'\in\mathcal{D}} \alpha_{t-d}(i,d') a_{ij} p_j(d) b_j(\boldsymbol{o}_{t-d+1:t})\end{aligned}$$

$$\begin{aligned}\beta_t(j) &\equiv P(\boldsymbol{o}_{t+1:T}|S_{t]} = j, \theta) \\ &= \textstyle\sum_{i\in\mathcal{S}\setminus\{j\}} \sum_{d'\in\mathcal{D}} a_{ji} p_i(d') b_i(\boldsymbol{o}_{t+1:t+d'}) \beta_{t+d'}(i)\end{aligned}$$

Here, a_{ij} is the state transition probability, i.e. the probability of transitioning from state i to state j, and $\mathcal{S}$ and $\mathcal{D}$ denote the set of all possible states and durations, respectively. The notation $S_{t]} = j$ indicates that layer j ends at t, and similarly we will use $S_{[t} = j$ to imply that layer j starts at t. For application to annual layer counting, several simplifications can be made. Most importantly, the Markov chain describing the changes in state along the data series is no longer a proper Markov chain, since layers are simply enumerated consecutively down the core without skipping. It follows that:

$$a_{ij} \equiv P(S_{[t+1} = j \mid S_{t]} = i) = \begin{cases} 1, & j = i+1 \\ 0, & otherwise \end{cases}$$

Further, layer duration is assumed to be independent of layer number, i.e. $p_j(d) = p(d)$. Here, $p(d)$ is taken to be a log-normal distribution (Andersen et al. 2006), however the subsequent equations do not depend on this choice. Introducing the notation $\bar{\alpha}_t(j) \equiv \sum_{d'\in\mathcal{D}} \alpha_t(j,d')$, the forward and backward equations can be reduced to:

$$\alpha_t(j,d) = p(d) b(\boldsymbol{o}_{t-d+1:t}) \bar{\alpha}_{t-d}(j-1) \qquad \text{(Eq. 3)}$$

$$\beta_t(j) = \textstyle\sum_{d'\in\mathcal{D}} p(d')\, b(\boldsymbol{o}_{t+1:t+d'})\, \beta_{t+d'}(j+1) \qquad \text{(Eq. 4)}$$

The following entities, which include the desired probabilities $\tilde{\gamma}_t(j)$ (Eq. 1), can now be calculated:

$$\eta_t(j,d) \equiv P(S_{[t-d+1:t]} = j, \boldsymbol{o}_{1:T}|\theta) = \alpha_t(j,d)\beta_t(j) \qquad \text{(Eq. 5)}$$

$$\begin{aligned}\gamma_t(j) &\equiv P(S_t = j, \boldsymbol{o}_{1:T}|\theta) \\ &= \gamma_{t+1}(j) + P(S_{t]} = j, \boldsymbol{o}_{1:T}|\theta) - P(S_{[t+1} = j, \boldsymbol{o}_{1:T}|\theta) \\ &= \gamma_{t+1}(j) + \textstyle\sum_{d\in\mathcal{D}} \eta_t(j,d) - \sum_{d\in\mathcal{D}} \eta_t(j-1,d)\end{aligned} \qquad \text{(Eq. 6)}$$

$$\tilde{\gamma}_t(j) \equiv P(S_t = j|\boldsymbol{o}_{1:T}, \theta) = \gamma_t(j)/P(\boldsymbol{o}_{1:T}|\theta) \qquad \text{(Eq. 7)}$$

$$P(\boldsymbol{o}_{1:T}|\theta) = \textstyle\sum_{j\in\mathcal{S}} P(S_t = j, \boldsymbol{o}_{1:T}|\theta) = \sum_{j\in\mathcal{S}} \gamma_t(j)$$

The equations (2-7) are implemented in log-space to prevent underflow.

To allow for comparison with manual counts, a constrained version of the Viterbi algorithm is subsequently used to translate the FB age distributions, $\tilde{\gamma}_t(j)$, into a consistent most likely set of layer boundaries (Winstrup 2011). The Viterbi algorithm uses the same model parameters, and is constrained to find the same total number of layers as the FB algorithm. *StratiCounter* output comprises the full FB age probability distributions as well as the most likely set of layer boundaries.

2.1. Boundary Conditions

To compute the forward and backward variables (Eq. 3-4), boundary conditions are required for the state of the system before and after the observed data.

As initial condition for the forward variable, the observed data always starts in layer 1, i.e. $S_{t=1} = 1$. Starting location of this first layer is given as a probability distribution. We have no knowledge about data values outside the available observations, and thus $b(\boldsymbol{o}_{\tau-d+1:\tau}) = 1$ for $\tau < 1$. The boundary condition for $j = 1$ reads:

$$\bar{\alpha}_\tau(0) = P(S_{\tau]} = 0|\theta),\ \tau < 1 \qquad \text{(Eq. 8)}$$

As boundary condition for the backward variable, the user can choose one of the following two options: The general assumption is no prior knowledge on the total number of layers, i.e. state of the system at and after T. This leads to:

$$\forall j: \quad \beta_\tau(j) = 1, \qquad \tau \geq T$$

This boundary condition is independent of j, hence $\beta_t(j)$ will be independent of j for all t.

Alternatively, the algorithm can be deployed with age constraints. Some volcanic eruptions emit large amounts of sulfuric acids, which can be transported very far from the eruption site. At the poles, the volcanic sulfur is deposited with the snow, thereby creating a distinct marker horizon in the ice core data. For recent times, historical eruptions produce depth-age marker horizons that can help validate the output of the algorithm and/or constrain its operation. Assuming that a data section starts and ends at depths corresponding to such age markers, the total number of layers in the section is known. Denoting this number J, the corresponding boundary condition reads:

$$\beta_\tau(j) = \begin{cases} 1, & j = J \\ 0, & otherwise \end{cases}, \qquad \tau \geq T$$

2.2. Statistical Description of an Annual Layer

Annual layers tend to display large variability in shape, which stems from a multitude of factors: Input of impurities to the atmosphere varies significantly from year to year, as do the large-scale weather patterns responsible for their transport to Greenland and Antarctica. The majority of impurities are deposited with the snow, causing the annual signals to also depend on the timing of the snowfall events. Finally, extreme events, such as volcanic eruptions, may occasionally overprint the annual signal.

To make the algorithm easily applicable to a wide range of chemistry series, *StratiCounter* first extracts a layer template based on rough manual layer counts of a section of the core. For each chemistry series, the layer template consists of a mean shape, $m(t')$, as well as the 1st principal component of the residuals from the mean shape, $r(t')$. For each chemistry series, a layer is modelled as:

$$y(t') = m(t') + A \cdot r(t') + \varepsilon(t'), \qquad 0 \leq t' < 1$$

We use a hierarchical approach to increase the flexibility of this layer description: The value of A is allowed to differ from year to year, but assumed to belong to a normal distribution, $A \sim N(A_0, \Phi)$. The error component $\varepsilon(t')$ is assumed white noise with $\varepsilon(t') \sim N(0, \sigma_n^2)$. Under these conditions, the likelihood that a given segment of a chemistry series exactly constitutes an annual layer can be computed using Bayesian linear regression (Bishop 2006). Assuming each chemistry series to be independent, the corresponding b-value (Eq. 2) is found by multiplication of their likelihoods.

To improve the layer description, derivatives of the chemistry series are included as additional data series. Layer parameters are tied between chemistry series and their derivatives, and are based on both. As the two data series are not independent, their layer likelihoods are combined to a single mean value. By including derivatives, information on the location of peaks and troughs within a layer is emphasized. Furthermore, the error structure becomes less rigid, since the error components on both data and derivatives are modelled as white noise.

2.3. Estimating Model Parameters

To describe the annual layers in the full data set of n chemistry series, we have n templates (each consisting of two shapes) and five parameters: two parameters describing the log-normal layer thickness distribution (μ_d and σ_d) and three parameters describing the expression of a layer: A_0 (nx1 vector), Φ (nx1 vector), and σ_n (nx1 vector). The templates, which provide the basic layer shapes, are held constant throughout the run of the algorithm, whereas the parameters are allowed to vary slowly with depth.

We use the Expectation-Maximization (EM) algorithm to optimize the parameter values based on the outcome of the FB algorithm. In each iteration, an optimized set of Maximum-Likelihood (ML) parameter values are computed (Winstrup 2011; Winstrup et al. 2012). The FB algorithm is run iteratively with these as parameters until convergence, or until a maximum iteration number is reached.

2.4. Implementation in Batches

The algorithm is run batch-wise down the chemistry series, with the length of each batch adjusted to contain approximately some predefined number of layers (default is 50) and having a slight overlap (up to 10%) with the previous batch. The quoted value of ~50 layers was chosen to provide a reasonable amount of data for the EM-optimization routine to converge.

Some information on previous layer locations is passed from one batch to the next. Upon completion of layer inference in a batch, an appropriate starting point for the next batch is selected immediately after the most unequivocal layer boundary (high values of $\sum_{j,d} \eta_t(j,d)$) near the end of the batch. The variable $\eta_t(j,d)$ also supplies the boundary condition for probable start locations of the first layer in the new batch (Eq. 8). In this way, the next batch is partly informed by data in the current batch, thereby improving the inference of layers in the overlap section.

For the first batch, an initial set of parameters is found based on a rough set of manual layer counts. For subsequent batches, ML estimates from the previous batch are used as starting point for the iterations of ML parameters for the current batch. Due to stability in layer

characteristics, convergence usually occurs after a few iterations (<4).

This procedure allows the layer characteristics to slowly evolve with depth, as they do in a real core. As an example, layer thicknesses in an ice core generally decrease with depth due to ice flow (Fig. 1A), so inferring layers in the entire core at once would violate the assumption that the layer thickness distribution is static. Similarly, other aspects of the annual layer signature may change over time, i.e. with depth. Hence, the batch size must be chosen to provide the best trade-off between containing a sufficient number of layers to provide reasonable estimates of the model parameters, while also being sufficiently short that the layer characteristics can be assumed constant within each batch.

3. Performance Evaluation

It is not straight-forward to evaluate the performance of the algorithm. The only direct evaluation available is comparison to a set of manual counts, but these are themselves associated with some uncertainty. A direct measure of their uncertainty based on inter-annotator agreement is usually unattainable; Standard practice is to focus on a reconciled set of layer assignments, effectively causing the initial sets of annual layers independently labeled by multiple experts to no-longer be available for analysis. Layer marks may also not be placed very accurately within a year, since this is generally not the main goal in manual layer counting. An objective comparison of the performance of the algorithm relative to manual counts is further complicated by the fact that the latter may have been informed by additional information, such as eruption ages for historical volcanoes. While it is possible also to incorporate such information into *StratiCounter*, this has not been done in the experiments described here.

We evaluate the similarity to a set of manual counts by a layer-to-layer comparison, inferring false-discovery rates and miss rates. Given that the precise location of layer boundaries might differ, we evaluate the number of inferred layer boundaries within a running weighted window, and note areas of discrepancy. A measure for the overall dissimilarity, D, is calculated as the square-root of the sum of the squared false-discovery and miss rates.

To obtain a baseline against which the obtained false-discovery and miss-rates can be compared, we generated 10,000 sets of random layer markers adhering to the same log-normal distribution as the manual layer assignments. The random layer markers provided false-discovery and miss rates that were approximately normally distributed with mean 5.7% and standard deviation 0.5%; the corresponding D-values were normal distributed with mean 8.0% and standard deviation 0.5%.

Note that this performance evaluation is based on the derived layer boundaries, and not the full probability output of the FB algorithm. While the most likely set of layer boundaries may differ from the manual layer counts, they may agree within the derived confidence interval. This is not accounted for here, and the calculated D-values may hence overestimate the actual differences.

In some cases, known ages of volcanic eruptions may serve as a second check on the obtained layering. While they do not allow for individual layer comparisons, they do provide a powerful check on the obtained ages at depths corresponding to volcanic eruptions. Yet this comparison can be difficult since volcanic eruptions happen frequently, and it may not be possible to tell them apart based on the data. Annotation of a sulfur spike to a specific eruption is often based on either a match to other layer-counted cores or directly on the layer-counting results. Further, it may take the volcanic sulfur a few months to a few years to reach the polar regions. Consequently, the sulfur increase observed in the ice core data may be delayed by up to a couple of years relative to the eruption.

4. Data

We use chemistry data from two ice cores drilled respectively in Greenland (NEEM-2011-S1, hereafter NEEMS1) and Antarctica (WAIS Divide Core, hereafter WDC), and focus on the period 1258-1815 AD. This section is demarcated by elevated sulfur levels caused by two large eruptions. Data are available online, along with a reliable set of manual layer counts, with the NEEMS1 layer counts tied to GICC05 (Sigl et al. 2013; Sigl et al. 2015b). For the selected period, known ages of large historical volcanic eruptions reduce uncertainties on the manual timescales for the two cores. The manual timescale uncertainties are therefore considered negligible.

We excluded chemistry series that replicated the annual signal in other data series, and ended up with seven different chemical series both for NEEMS1 and WDC. The employed chemistry series were: black carbon, Br (WDC only), HNO_3 (NEEMS1 only), NH_4, Na, non-sea-salt Ca, non-sea-salt S, and non-sea-salt-S-to-Na ratio. In the format provided online, long-term trends had been removed and data was normalized. To enhance the layer signal additionally, data was preprocessed by computing z-scores over a 1m running window, and averaged to 2cm resolution.

5. Experiments

5.1. Dependency on Input Data

We first ran *StratiCounter* for NEEMS1 and WDC, while varying the number and type of chemical series used for timescale inference, and recorded false-discovery rates and miss rates. These rates are dependent on the performance of the algorithm, as well as on the predictive ability of the employed chemistry series(s).

Various chemistry series convey different information, and display different types of errors. All *StratiCounter* outputs were distinctly different from a random distribution with correct layer thickness distribution. When using a single chemistry series to obtain annual layering, we observed a tendency towards counting too many layers rather than too few: one chemical series resulted in a timescale with a false-discovery rate as high as 33%, whereas the maximum miss rate was 7.3%. Minimum values of false-discovery and miss rates obtained for single chemistry series were 1.3% and 0.3%, respectively, and minimum D-values were 2.1% for NEEMS1 and 12.1% for WDC. However, the same chemistry series did not work equally well at the two ice core locations, so no general conclusion can be drawn regarding the best chemistry series to use for annual layer dating.

To explore the added value of incorporating multiple data series, we did an exhaustive search of the layering obtained when combining an increasing number of chemistry series. Overall, incorporating more chemical series resulted in increased accuracy, although the amount depended on the selected chemistry series. For NEEMS1, the median of the D-values showed rapid decrease when including the first extra chemistry series (from 14% (one series) to 3.2% (three series)), but with diminishing returns after including more than four chemistry series. The same tendency was observed for WDC, although the values were a little higher overall (from 18% (one series) to 5.0% (three series)). For WDC, best agreement with manual counts was obtained when using all seven chemistry series. For NEEMS1, the best agreement was achieved when using three chemistry series (Na, non-sea-salt S, and non-sea-salt-S-to-Na ratio). These three series were among the primary ones employed in the manual approach (Sigl et al. 2013), this possibly causing the high similarity of the resulting timescales.

5.2. Evaluation of the All-Chemistry Timescales

For the timescales constructed with the full set of seven chemical series, all batches had converged within less than four iterations. The false-discovery and miss rates for NEEMS1 were 0.98% and 1.6% (D = 1.9%), and 1.2% and 1.2% (D = 1.6%) for WDC, much smaller than obtained for any of the chemistry series individually. These values are within the uncertainty range of manual layer counts for high-quality data, when not including age information from volcanic eruptions.

To further evaluate the timescales for the two cores, we considered the derived ages at the depth coinciding with a volcanic sulfur peak previously estimated by manual layer counting to occur in 1258 AD (Sigl et al. 2015b). Since the beginning of our selected depth interval was demarcated by sulfur from the 1815 AD Tambora eruption, this allows a check on the accumulated number of layers in the interval. The obtained 95% confidence intervals are 1258-1263 AD for NEEMS1, and 1255-1262 AD for WDC, with the most likely ages being 1261 and 1258 AD, respectively. For both cores, the ML age estimates are 3 years or less away from that obtained manually, and the manual estimate is covered by the associated confidence intervals.

6. Future Work

The two data sets employed here are nearly ideal for annual layer counting. *StratiCounter* also performs well for data with less-distinct annual layers (Winstrup et al. 2012; Vallelonga et al. 2014; Sigl et al. 2015a). Difficult data, however, tend to confound both manual and automated approaches alike, making it hard to evaluate performance of the algorithm by comparing to manual counts.

Given their importance for correctly identifying an annual layer, layer templates and associated layer likelihoods are main focus points for further development of the algorithm. Currently, the layer templates are held constant for the entire core. This is likely not a reasonable assumption when dealing with longer core sections, where layers may have formed under different climate conditions. We are working on introducing more flexibility in the layer template and layer likelihood computations, which will allow these to better mimic the processes responsible for creating the layer signature in the data.

To a large degree, manual layer counting is performed in sections step-wise down the core, as in *StratiCounter*. However, a human expert will often go back and forth between sections, matching the appearance of layers across them. In contrast, the current implementation of *StratiCounter* optimizes the model parameters describing layer characteristics for each batch individually. We can to some degree imitate the manual passing of information between batches by introducing priors on the model parameters. With priors based on layers in previous batches, model parameters in adjacent batches will be correlated, and layer characteristics will be forced to change more slowly with depth. This feature is under development.

Additionally, the calculation of confidence intervals must be revised to include uncertainty on the employed

model parameters. When estimating uncertainty bounds, the algorithm currently assumes known model parameters, and consequently these bounds are too narrow. This is particularly evident when running *StratiCounter* on data that do not contain sufficient information to produce good estimates for the model parameters. By including parameter uncertainty, we expect to achieve layer-counted timescales with more reliable confidence intervals.

The *StratiCounter* code can be downloaded from: www.github.com/maiwinstrup/straticounter.

Acknowledgments

This work was supported by grants from the Villum Foundation and the Inge Lehmann Foundation. I thank Elizabeth Bradley for scientific guidance and thoughtful discussions. Many thanks also to Neil Jacobstein and two anonymous reviewers for constructive comments that improved the manuscript. I gratefully acknowledge the efforts of the many people involved in the logistics, drilling, ice-core processing and chemical analyses of the NEEMS1 and WAIS Divide ice cores. The WAIS Divide ice core project is funded by the US National Science Foundation (US NSF). NEEM is directed and organized by the Centre of Ice and Climate at the Niels Bohr Institute, University of Copenhagen and US NSF, Office of Polar Programs. It is supported by funding agencies and institutions in Belgium (FNRS-CFB and FWO), Canada (NRCan/GSC), China (CAS), Denmark (FIST), France (IPEV, CNRS/INSU, CEA and ANR), Germany (AWI), Iceland (RannIs), Japan (NIPR), South Korea (KOPRI), The Netherlands (NWO/ ALW), Sweden (VR), Switzerland (SNF), the United Kingdom (NERC), USA (US NSF, Office of Polar Programs), and the EU Seventh Framework program (Past4Future, WaterUnderTheIce).

References

Andersen, K.K. et al., 2006. The Greenland Ice Core Chronology 2005, 15–42ka. Part 1: constructing the time scale. *Quaternary Science Reviews* 25: 3246–3257.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*, Springer Science + Business Media, LLC.

Buizert, C. et al., 2015. The WAIS Divide deep ice core WD2014 chronology - Part 1: Methane synchronization (68–31 ka BP) and the gas age–ice age difference. *Climate of the Past* 11: 153–173.

Hammer, C. et al., 1978. Dating of Greenland ice cores by flow models, isotopes, volcanic debris, and continental dust. *Journal of Glaciology* 20(82): 3–26.

Jouzel, J. et al., 2007. Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* 317: 793–796.

McGwire, K.C. et al., 2011. Instruments and Methods: Identifying annual peaks in dielectric profiles with a selection curve. *Journal of Glaciology* 57(204): 763–769.

NEEM community members, 2013. Eemian interglacial reconstructed from a Greenland folded ice core. *Nature* 493: 489–494.

Rasmussen, S.O. et al., 2006. A new Greenland ice core chronology for the last glacial termination. *Journal of Geophysical Research* 111(D6): D06102.

Rasmussen, S.O., Svensson, A.M., and Winstrup, M., 2014. State of the art of ice core annual layer dating. *Past Global Changes Magazine* 22(1): 26–27.

Sigl, M. et al., 2013. A new bipolar ice core record of volcanism from WAIS Divide and NEEM and implications for climate forcing of the last 2000 years. *Journal of Geophysical Research: Atmospheres* 118: 1151–1169.

Sigl, M. et al., 2015a. The WAIS Divide deep ice core WD2014 chronology – Part 2: Annual-layer counting (0–31 ka BP). *Climate of the Past Discussions* 11: 3425–3474.

Sigl, M. et al., 2015b. Timing and climate forcing of volcanic eruptions for the past 2,500 years. *Nature* 523: 543–549.

Steffensen, J.P. et al., 2007. High-Resolution Greenland Ice Core Data Show Abrupt Climate Change Happens in Few Years. *Science* 321: 680–684.

Svensson, A.M. et al., 2008. A 60,000 year Greenland stratigraphic ice core chronology. *Climate of the Past* 4: 47–57.

Vallelonga, P. et al., 2014. Initial Results from Geophysical Surveys and Shallow Coring of the Northeast Greenland Ice Stream (NEGIS). *The Cryosphere* 8: 1275–87.

Winstrup, M. 2011. An Automated Method for Annual Layer Counting in Ice Cores - and an Application to Visual Stratigraphy from the NGRIP Ice Core. Ph.D. diss., Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

Wheatley, J.J. et al., 2014. Bayesian Layer Counting in Ice-Cores: Reconstructing the Time Scale. In *The Contribution of Young Researchers to Bayesian Statistics,* Springer Proceedings in Mathematics & Statistics: 121–125.

Winstrup, M. et al., 2012. An automated approach for annual layer counting in ice cores. *Climate of the Past* 8: 1881–1895.

Yu, S.Z., 2010. Hidden semi-Markov models. *Artificial Intelligence* 174: 215–243.