FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

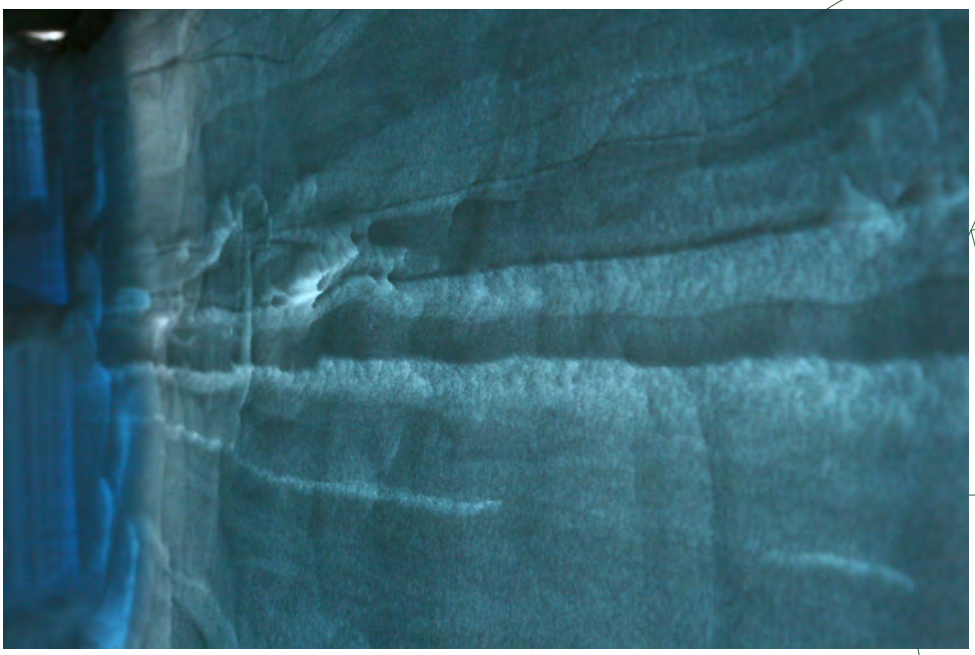
PhD Thesis
Mai Winstrup

An Automated Method for Annual Layer Counting in Ice Cores

- and an application to visual stratigraphy data from the NGRIP ice core

Academic supervisor: Anders Svensson
Co-supervisor: Sune Olander Rasmussen

Submitted 23/10-2011



Visual stratigraphy in a double pit at NEEM, Summer 2009

CENTRE FOR ICE AND CLIMATE

Preface

The story about this thesis starts out during my stay at the University of Washington, Seattle. After work. Over a beer. And me being very tired of everything to do with layer detection models that always appear to be somewhat working, but aren't really. And always because of the same: They lack the ability to see what comes after the next layer. "Why don't you use a Hidden Markov Model?" Amittai asks me. "You know, the Forward-Backward algorithm. It ought to do the job." And I'm looking completely ignorant. But "Forward-Backward" sounds good. Forward pass. Backward pass. Using all observations at once.

And here it is, my thesis. All wrapped up in Hidden Markov Models and Forward-Backward algorithms. I certainly didn't have that in mind, when I started out on my PhD.

Thank you, Amittai Axelrod, for introducing me to Hidden Markov Models! You certainly gave me a lot of work, but without you, my thesis would have been completely different.

I wish to thank Anders Svensson for being my supervisor, and for letting me do this. Lots of thanks to Sune Olander Rasmussen for support and awesome comments. Many thanks also to Ole Winther for helping out with the statistics whenever I got stuck. I also wish to thank Ed Waddington and Eric Steig for providing a lot of support during my stay at the University of Washington. And thanks to Jesper Sjolte, who suggested many improvements to the manuscript.

So many others deserve my thanks for having contributed to this project in some way or another, but as space is limited, I'll do it in person instead.

Thank you to everybody at CIC for being so great colleagues and friends.

And last, but not least: Lots of love to my family who is always there for me.

This thesis is submitted in partial fulfillment of the requirements for the PhD degree at Centre for Ice and Climate, Niels Bohr Institute, University of Copenhagen, Denmark.

Mai Winstrup,
Copenhagen, October 2011

Abstract

An accurate chronology is of fundamental importance for the interpretation of a paleoclimatic record. The high temporal resolution of the Greenland ice cores has allowed the construction of an annual layer counted chronology for these reaching back to 60 ka BP, the oldest part of which is based on the NGRIP ice core. But as the annual layers become thinner towards the bed, the annual signal in most components weakens, and the subjectivity involved in manual layer interpretation increases. To extend the layer counted chronology beyond 60 ka BP, a more objective methodology of layer detection is needed.

For this purpose, an automated layer detection algorithm has been developed. It is based on the statistical framework of Hidden Markov Models (HMMs), originally developed for use in speech recognition. Meticulously based on statistical considerations, the algorithm is able to determine the most likely annual layering in an entire data section at once. The fundamental strength of the algorithm lies in the way that it is able to imitate the manual procedures, while being based on purely objective criteria for annual layer recognition.

The algorithm has been implemented for the visual stratigraphy data from NGRIP, in which the annual signal is covered in noise, but maintained to great depths. The algorithm is tested for three sections: A cold period (GS-13), a warm period (GI-12), and the transition between the two. The algorithm has not yet been tuned to provide an accurate chronology, but the results look promising. The algorithm was e.g. able to obtain a good result when passing over the transition period with a corresponding halving in annual layer thicknesses over merely five meters.

Resumé

Det er vigtigt for fortolkningen af en palæoklimatisk tidsserie også at have en nøjagtig tidsskala. Den høje tidslige opløsning af de grønlandske iskerner har gjort det muligt for disse at danne en årlagsoptalt tidsskala, der går 60.000 år tilbage. Den ældste del af denne er dannet på baggrund af data fra NGRIP iskernen. Men efterhånden som årlagene bliver tyndere nedefter i iskernen, bliver årlagssignalet i de fleste komponenter svagere, og den manuelle optælling bliver mere og mere subjektiv. Det er derfor ikke muligt manuelt at føre tidsskalaen længere tilbage i tiden.

Til dette formål er der blevet udviklet en algoritme, der automatisk kan finde årlagene. Algoritmen er baseret på et statistisk grundlag kaldet Hidden Markov Model (HMM), som oprindeligt er blevet udviklet til talegenkendelse. Baseret på grundige statistiske overvejelser er algoritmen i stand til at finde de mest sandsynlige årlag i en hel datasektion på een gang. Algoritmens styrke ligger i den måde, hvorpå den er i stand til at efterligne en manuel tilgang til årlagsgenkendelse, men samtidig er baseret på objektive kriterier.

Algoritmen er udviklet til brug på data fra den visuelle stratigrafi fra NGRIP. I denne dataserie er det årlige signal ganske vist meget påvirket af støj, men det har forblevet intakt ned til en stor dybde. Algoritmen er afprøvet for 3 sektioner: En kold periode (GS-13), en varm periode (GI-12), og overgangen mellem de to. Algoritmen er endnu ikke færdigudviklet, men resultaterne er lovende. Algoritmen var f.eks. i stand til at opnå et godt resultat for overgangen mellem den kolde og den varme periode, hvor der skete en halvering af årlagstykkelserne over blot fem meter.

Contents

1. Introduction

The value of a paleoclimatic record ultimately depends on the acquired knowledge on its associated timescale. A timescale is needed for comparing different paleoclimatic proxies as well as for answering questions on e.g. periodicities and rapidity of shifts in the climate system, both of which may help to improve our knowledge on the involved climatic processes. It is therefore a general challenge, relevant to all climatic proxies, to obtain a timescale as accurate as possible.

Among the most precise are chronologies based on paleoclimatic archives containing annually laminated data. Such archives include tree rings, varves (seasonally laminated organic lake deposits), corals – and ice cores. The subannual resolution of these records provides an opportunity to count annual layers back in time. The ease with which such counting can be carried out, and hence the accuracy of the resulting chronology, depends on the data record in question. In this respect, dendrochronology (tree ring counting) is probably the most famous, possibly providing an almost perfect chronology several thousand years back. Greenland ice cores follow right after. The Greenland Ice Core Chronology 2005 (GICC05) is an annually counted chronology based on a composite of Greenland ice cores, which goes back to 60.000 years BP with an estimated total uncertainty of 2600 years.

In a textbook on dendrochronology, it has once been stated that “two or three cores should be taken from each tree and at least 20-30 trees sampled at an individual site” [*Bradley*, 1985, p. 334-335]. In this way, questions arising on layers which are difficult to interpret can to a large extend be resolved. When dealing with ice cores, practical difficulties and costs associated with core recovery does not allow for such practice to take place. In its place, a multi-parameter method can be applied. A range of chemical impurities, as well as the stable water isotopes, display a seasonal signal, and by combining the information in as many of these as possible, an accurate chronology can be achieved as far down as data quality allows.

In this chapter, I will first outline the different methodologies which are used for dating the Greenland ice cores. I will then touch on why the NGRIP ice core provides an exceptional opportunity to construct a high-resolution layer counted chronology far back in time, and describe in more details how this annual layer counting was carried out. Howev-

er, the subjectivity involved in layer interpretation is increasing with depth, and to extend the GICC05 beyond 60 ka BP, a more objective methodology of layer detection is needed. The remaining part of the thesis describes the development of a statistical framework which in the future might be used for such purpose.

1.1 Greenland ice core timescales

Many different approaches can be used for dating ice cores, the applicability of each depending on the specific situation and the amount of data available. The most precise timescales are stratigraphically based. However, if data for constructing such timescales is not available, more or less elaborate ice flow models can be used to estimate the age-depth relationship.

1.1.1 Modeled timescales

Knowledge derived from ice flow models on stress and strain rates in the ice sheet can be used for predicting the rate at which annual layers are thinning with depth, and hence for constructing a timescale for an ice core. The employed ice flow models span from simple 1D models [*Dansgaard and Johnsen,* 1969] to much more elaborate ones [*Parrenin et al.*, 2004].

In advance of obtaining an ice core, modeled timescales constructed by ice flow models may be used for selecting the best location for the ice core to be drilled [*Dahl-Jensen et al.*, 1997]. But after retrieval of the core, information gained from the ice core data may be incorporated into the model. Studies have e.g. shown a strong correlation between past accumulation rates and the relative concentration of stable water isotopes ($\delta^{18}O$) in the ice core [*D Dahl-Jensen et al.*, 1993], and such information may be incorporated into the model. Ice flow models may also be combined with age markers found from the ice core data. This combination has been used for establishing timescales for several of the Antarctic ice cores [*Parrenin et al.*, 2001; *Parrenin et al.*, 2007].

The Dansgaard-Johnsen model is a simple ice flow model commonly used to provide timescales for the Greenland ice cores. In this model, the vertical strain rates as a function of depth are derived based on mass conservation and a predefined horizontal velocity profile [*Dansgaard and Johnsen*, 1969]. Despite its simplicity, and in so far that the adjacent flow regime is relatively simple, it often yields quite satisfying results. It provides the basis for the ss09sea timescale [*S J Johnsen et al.*, 2001], which originally was constructed for the GRIP ice core, and later modified to account for basal melting and applied to the NGRIP ice core (see map in figure 1.3.1). The model integrates knowledge on past accumulation rates, and the resulting timescale turned out to be in good agreement with the later constructed annual layer counted chronology GICC05 [*Svensson et al.*, 2006]. The model can also be used to convert from annual layer thicknesses in a given depth to an estimate of past accumulation rates, a relationship referred to in section 4.1.

In some cases, a modeled time scale is indeed the only option for obtaining a timescale for an ice core. This is almost always the case for the lower part of an ice core, where annual layers are thin and difficult to identify. In Antarctica, the often very low accumulation

rates to a high degree eliminate the possibilities of stratigraphic dating. At such places, dating by ice flow modeling is one of the only means for producing a timescale for the ice cores.

1.1.2 Stratigraphic dating of ice cores

Stratigraphic dating of ice cores covers both the use of reference horizons to link particular features in the ice core record to a fixed chronology, and the use of annually resolved data to count annual layers.

Numerous events can create reference horizons in the ice core. All that is needed is the event to somehow stand out in the ice core data. The reference horizons most commonly used are layers of high concentrations of sulphuric acid, which often are related to volcanic events. For layers corresponding to well-known volcanic events during historical times, such layers can be used as fix points in the timescale [*Hammer*, 1980]. But even when the volcanic event has not been independently dated, such reference horizons can be used to link individual paleoclimatic records [*Vinther et al.*, 2006].

Due to the regularity with which snow is deposited on the ice sheet surface and gradually compressed into ice and thinned during ice flow, ice has the ability to preserve a very reliable climate record. Provided that the accumulation rate is sufficiently high, there is low risk of missing years in the record, and if indeed an annual layering has been preserved, it can be used to establish a counted chronology down the ice core.

Such counted chronologies have mainly been constructed for the upper part of ice cores, where annual layers not yet have been thinned too much [*Hammer et al.*, 1978; *S J Johnsen et al.*, 1992]. This methodology for providing a timescale is particularly useful at high-accumulation sites. The annual layers here maintain a reasonably large layer thickness for the longest time interval, thereby allowing these layers to be detectable. A frequently applied timescale is the Meese-Sowers timescale [*Alley et al.*, 1997; *Meese et al.*, 1997], a counted chronology established for the GISP-2 ice core from Central Greenland (see a further description in section 2.4).

Because of the way uncertainties are introduced in an annually counted chronology, such timescales provide good estimates on the relative timing of two events, whereas the absolute dating uncertainties may be rather large.

Other methods than those mentioned above can also be used for establishing a chronology for ice core records. This includes e.g. wiggle-matching to existing ice cores and/or ocean cores. And very often a timescale for an ice core is established based on a combination of all of the above. A very comprehensive example of this is found in *Lemiuex-Dudon* [2010], where flow modeling, age markers from several ice cores etc. are utilized to make a consistent timescale for several ice cores at a time.

1.2 Ice core data of subannual resolution

The isotopic composition and impurity content of snow deposited in the inner part of an ice sheet is depending on climate as well as time of the year. Both of these variations are

recorded in the deposited ice, where the signal may remain unchanged over long time periods. With the high temporal resolution of the Greenland ice cores in particular, these can therefore be used not only for inferring past changes in climate, but they may also allow seasonal information to be inferred.

To obtain an ice core record of subannual resolution as required for establishing a counted chronology, the chemical impurities in the ice must be measured in high resolution. The resolution necessary depends on the thickness of the annual layers in question. With depth, the annual layers are thinned, and higher demands are imposed on the measurement techniques.

High-resolution impurity profiles of ice cores can be measured using the method of Continuous Flow Analysis (CFA). The basic idea behind CFA-measurements is to continuously melt a rod of the ice core on a melt head. The melt stream from the inner uncontaminated part of the ice is then let through a multitude of analytical lines, each of which measures the concentration of a specific chemical component in the ice core melt water [*Rothlisberger et al.*, 2000].

The resolution of a set-up of the CFA system depends on the number of chemical species being measured, in combination with the melt speed and the mixing volumes of the sample stream when being transported through the system. By minimizing mixing volumes, reducing the number of components being measured, and using a slow melt speed, the chemical concentrations can be measured in very high details. Presently, the highest resolution obtained in a CFA system is able to resolve annual layers down to 1 cm in thickness [*Bigler et al.*, 2011], potentially allowing annual layers to be resolved in both Greenland and Antarctic ice cores throughout the last glacial cycle.

Whether or not annual layers can be resolved in ice core data does not only depend on the measurement technique. In the upper part of an ice core, a very clear annual signal is often observed in the $\delta^{18}O$-record of the ice, caused by the large temperature differences between summer and winter. With depth, this signal is slowly obliterated. A large degree of diffusion during the firnification process, but also molecular diffusion in the ice, causes the annual layering to slowly dissolve. To some extent, the annual signal can be reconstructed in the deeper ice using back-diffusion methods [*S Johnsen*, 1977], but at some point, the seasonal variations will have disappeared.

Similarly holds for the remaining chemistry in the ice core. With depth, diffusion slowly diminishes their annual signal, which at some point will have disappeared. Only a few records are not much affected by this process. These records are generally records which are somehow connected to the existence of larger particles, such as e.g. dust particles, for which diffusion generally is negligible.

1.3 The NGRIP ice core: A chronologist's delight

The North Greenland Ice core Project (NGRIP) took place in Northern Greenland (figure 1.3.1) during the years 1995-2003 as a joint international deep ice core drilling project. The NGRIP ice core is 3085 m long and goes back approximately 123.000 years to the

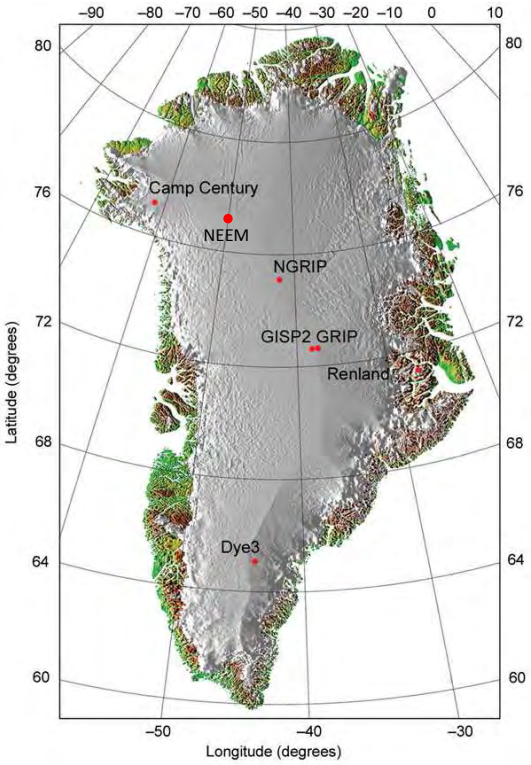


Figure 1.3.1: *Locations of Greenland deep ice core drill sites.*

start of the last interglacial, the Eemian (115-130 ka BP) [*D Dahl-Jensen et al.*, 2002; *North Greenland Ice Core Project Members*, 2004]. Presently, the accumulation rate at the location is 0.195 cm ice equivalent per year, and the mean annual temperature is -32°C [*D Dahl-Jensen et al.*, 2002].

Originally, the main purpose of the NGRIP ice core was to retrieve a full and undisturbed sequence of ice from the last interglacial. The ice core records from two previous deep-drilling ventures located at the summit of the Greenland ice sheet (GRIP and GISP-2) both contain Eemian ice. However, the stratigraphy in the lower part of the two cores turned out to be dissimilar, and both of the records were folded at depth, probably due to ice flow over a bedrock topography known to have relatively large undulations [*D Dahl-Jensen et al.*, 2002]. For this reason, the degree of climate variability in Greenland during the Eemian was still a puzzle, and the NGRIP drill site was selected to yield an answer to exactly that: From radio-echo sounding (RES), the bedrock was found to be flat in the area, and the internal layering in the ice seemed to promise a good spot for retrieving Eemian ice [*D Dahl-Jensen et al.*, 2002].

Yet, at the NGRIP site, the existence of unexpectedly high basal melt rates in the area turned out to have a major effect on the stratigraphy of the ice core. Basal melting had simply removed the oldest ice, and the ice core only reached into the first part of the Eemian. Although disappointing at first, the high basal melt rates turned out to provide a unique possibility for developing an accurate ice core chronology for the Greenland ice cores.

In the upper part of the NGRIP ice core, the relatively small accumulation rate cause the annual layers to be relatively thin in comparison to e.g. the Dye-3 ice core in Southern Greenland. But the high basal melt rates result in a much lower thinning rate of the annual layers, giving rise to a high time-resolution at depth. In combination with a concurrent development of sensitive and high-resolution CFA measurement techniques, it was possible to distinguish seasonal variation in the different chemical species even at great depths. Consequently, the NGRIP ice core provides the optimal conditions for carrying out high-resolution analyses, and hence to resolve annual layers in the ice core data.

1.4 Development of GICC05

Due to their high temporal resolution, Greenland ice cores can be dated very precisely by annual layer counting. A several year-long effort of manual annual layer counting using multiple chemical components has resulted in the Greenland Ice Core Chronology (GICC05), a composite and independent chronology common to three Greenland ice cores: Dye-3, GRIP and NGRIP. At each depth interval, the ice core with the highest resolution was used to count the annual layering, and the separate sections of timescales were subsequently pieced together using marker horizons [*Vinther et al.*, 2006]. The oldest part of the chronology is exclusively based on data from the NGRIP ice core.

1.4.1 A multi-parameter approach

The chronology is based on a multi-parameter approach, which make use of the range of high-resolution data sets available: Electrical Conductivity Measurements (ECM) of the solid ice [*Hammer*, 1980], melt water conductivity, concentrations of the impurities Na^+, Ca^{2+}, SO_4^{2-}, NO_3^-, NH_4^+, and the visual stratigraphy of the ice core (see chapter 2). In the upper part of the ice core, also the seasonal variation in the stable water isotopes $\delta^{18}O$ was used. In sections of data loss, layer boundaries were interpolated based on the layer thicknesses above and below.

The above mentioned components represent a diversity of environments. Ammonium (NH_4^+) is e.g. related to biological processes and biomass burning, while sodium (Na^+) to a first order has a marine source. The electrolytic conductivity of the melt water is a bulk signal of all the ionic constituents in the ice. These chemical components have different patterns of seasonality, and are affected differently by non-seasonal events.

However, even with many data records, the layering is sometimes ambiguous. After deposition, the seasonality of the signal can be altered due to e.g. melt events, wind scouring or snow drifting, thereby disrupting the signal in the stratigraphy. In regions of low accumulation rates, this may give rise to missing layers. Fortunately, accumulation rates are too high at the NGRIP site for this to have a major impact [*Andersen et al.*, 2006b], and periods of melt does not happen very often. Furthermore, non-annual features may exist, which may mistakenly be interpreted as annuals, or alternatively, they may obscure an underlying annual signal. In general, the annual signal is quite variable in most of the data series, hence complicating the interpretation, and emphasizing the need to use more than just a single data series.

The inclusion of many data series, which peak at different times of the year, and are affected differently by events of non-annual nature, such as e.g. volcanic events, results in a much more robust counting approach than if based on a single data series alone. By continuously examining the evolution in the annual layer signal, and learning their signal characteristics, the annual layers could be recognized with high certainty. Within the warm periods, the different species were peaking at different times during the year, and this knowledge was used when counting the annual layers. During the cold periods, the species were observed to peak more or less simultaneously [*Andersen et al.*, 2006b].

However, due to a decrease in annual layer thickness, it is not possible to continue the annual layer counting further back than 60 kyr BP based on the CFA multi-parameter data sets. Diffusion of the chemical species in the ice core as well as during the measurement, combined with an annual layer thickness dropping below 1 cm at this depth, effectively obliterates the annual signal in most of the ice core chemistry data [*Svensson et al.*, 2008]. Only the visual stratigraphy of the ice core maintains an annual signal in the deepest part of the ice core [*Svensson et al.*, 2005].

1.4.2 Uncertainty of GICC05

Given that some layers were ambiguous, the GICC05 chronology was made with an uncertainty estimate. When encountering an ambiguous layer (judged to be between 25-75% certain), such layer was counted as ½±½ year [*Rasmussen et al.*, 2006]. In this way, an estimate of the maximum counting error (abbreviated MCE) was produced. The MCE may be regarded as the 2σ-uncertainty band [*Andersen et al.*, 2006b]. This is a conservative uncertainty estimate, which acknowledges that the layer counting might be slightly biased, and it gives rise to an approximately linear increase in uncertainty with depth.

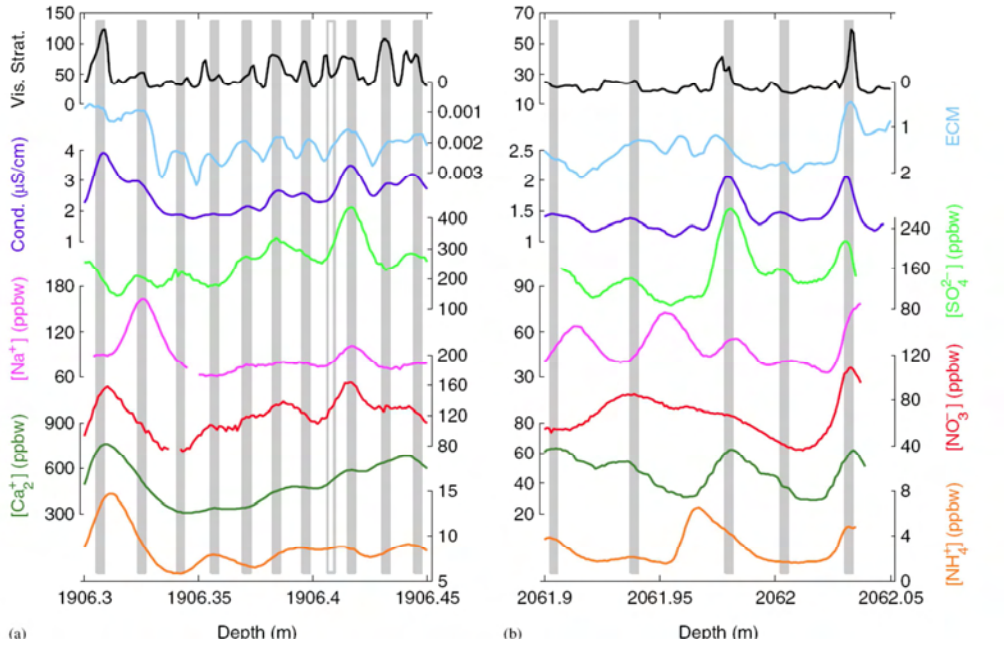


Figure 1.4.1: *Examples of the counting strategy. Grey bars are certain layers, white bars are considered uncertain layers and counted as ½ ±½ year. Figure reproduced from Andersen et al. [2006b].*

The upper part of the GICC05 chronology could be reconstructed very precisely, and with an uncertainty of only 87 years at a depth of 1400m, approximately corresponding to the onset of the Holocene. Further down, the uncertainties increase to around 4% during the warm periods, and 7% during the cold periods [*Andersen et al.*, 2006b]. The difficulties of recognizing layers increased with depth, and at a depth of 2426m, corresponding to 60 ka BP, the uncertainty estimate on the chronology is 2601 years.

In an attempt to eliminate as much as possible the subjectivity involved in layer interpretation, the annual layers were always counted multiple times, and by at least two experienced investigators. At first, each investigator counted a section by him/herself, and subsequently the two counting outcomes were compared. In case of large differences, the counting was redone in collaboration between the two investigators to reach consensus.

As the annual layers get thinner towards the bed, the annual signal in most components weakens. Diffusion of the different chemical species with depth in the ice core slowly causes the number of parameters containing an annual signal to become fewer. With their decreasing annual layer thicknesses, decreased resolution due to mixing during the measurement line also became more influential. In sections with small layer thicknesses, such as during the stadials in the deep ice, only a few parameters were left which had maintained their annual layer signal. These were the visual stratigraphy, the conductivity and ECM. To a first order, all of these are related to the dust content in the ice core.

At depths below 2430m, corresponding to an age of 60 ka BP, the annual layer thicknesses reach below 1 cm. At this point, only the annual layering in the visual stratigraphy is still intact. However, the annual layer signal in this data series is difficult to identify: At some years no peak is present, while several peaks may occur during others. Furthermore, with only a single data series, the subjective interpretation of the layering sequence tends to become an influential factor. At this depth, objective annual layer counting was considered impossible, and it was therefore decided not to carry on with the counting.

However, it has proved possible to recognize annual layers further down the NGRIP ice core, where annual layer thicknesses again reach above the 1 cm limit. Annual layers have e.g. been identified during sections of 120 ka old Eemian ice from the deepest part of the NGRIP ice core [*Svensson et al.*, Submitted 2011]. Likewise, annual layers may also be distinguishable during the warmer periods of the last glacial.

1.4.3 The resulting timescale

In figure 1.4.2 the resulting GICC05 timescale is compared to other independently dated records (ice cores and cave records) covering the same time period. The gray band is the 1σ-uncertainty band on the GICC05 timescale. When considering smaller sections, the individual chronologies may differ, but the overall agreement is good. The cave records are absolute dated with high precision. The agreement between the absolute dated Hulu cave [*Wang et al.*, 2001] and the relative dated GICC05 throughout the depth interval signifies that neither has a significant dating error.

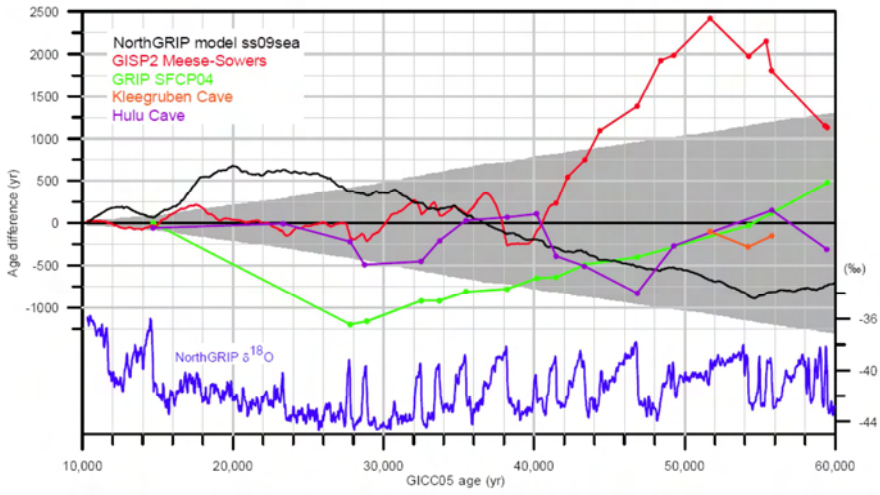


Figure 1.4.2: *Comparing GICC05 to other independently dated records: three ice core timescales (ss09sea, Meese-Sowers, SFCP04) and two cave records (Kleegruben, Hulu). The gray band is the 1σ-uncertainty band on the GICC05 timescale. Positive value means that the record is younger than GICC05. Figure reproduced from Svensson et al. [2008].*

1.5 Outline of thesis

In chapter 2, I will describe in more details the visual stratigraphy of an ice core as it has been measured on the NGRIP ice core, what these data are believed to be a record of, and describe the annual signal in this data series. Furthermore, it will be described how the data series here has gone through an extensive treatment to adjust for defects in the measurement device, something which turned out to be an absolute prerequisite for the further analysis.

The subsequent chapters deal with the development of a statistical framework that in an objective and robust manner is able to detect the annual layering in data where only a noisy annual layer signal is present. It has been developed with the NGRIP visual stratigraphy data in mind, but a similar method may also be useful for layer detection in many other kinds of data with annual laminations.

Chapter 3 deals with the development of an algorithm to detect the annual layering in a small data section, in which layers can be assumed uniform with respect to thickness and expression in the data series. This chapter provides the general probabilistic framework, which reduces the complex question of performing an overall pattern matching of multiple successive layers at a time, to the much simpler question of determining whether or not a particular data segment is likely to represent an annual layer. The development of a probabilistic description of an annual layer in the visual stratigraphy data is taken up in chapter 4.

A layer detection algorithm must be allowed to adapt to the constantly changing characteristic of an annual layer with depth. In chapter 5, equations are derived for a scheme of iterations, which enable the algorithm to adapt to such changes by using the data itself to make an appropriate choice of parameters describing e.g. the mean layer thickness. In

principle, this eliminates the need for including any knowledge based on previous data sections, and each batch of data can be processed independently. Hence, while maintaining the assumption of fixed parameter values within a batch, the layering in each of these is allowed to be described by its own set of parameters.

By excluding all information based on previous data, a direct implementation of the above mentioned iterative scheme is generally not very robust, especially not for data where the annual layering is disguised by many other types of variability. In chapter 5.4, it is therefore described how these iterations can be modified to take prior information on the appropriate parameter values into account.

Chapter 6 deals with the assemblage of sequential batches of data. It covers the practical aspects of joining the results from individual data sections, as well as a discussion on how the parameters employed here are expected to vary with depth. Also, it is outlined how the above mentioned iterative scheme can be taken yet another step further to allow the parameter values to continually be adjusted with depth in a proper sequential manner.

The theoretical part of the thesis ends in chapter 7 with a discussion on how the similarity between two independent annual layering sequences covering the same depth interval can be assessed. The here developed measures of similarity are later used for evaluating the performance of the algorithm.

In chapter 8, the results of a series of sensitivity studies of the algorithm are presented and discussed. Finally, in chapter 9, the algorithm has been employed on a representative section of the visual stratigraphy data from the NGRIP ice core. The selected section covers a warm period, the Greenland Interstadial 12 (GI-12) (depth: 2200-2220 m), and the preceding cold period (depth: 2225-2240 m). The associated time interval is approximately 45.9 to 48.3 ka BP. The inferred layering is compared to that of the GICC05 chronology. Also the performance of the algorithm over the transition from warm to cold period will be presented and discussed.

According to the GICC05 timescale, the entire section considered spans 2333 ± 121 years. This only represents a small part of the data available for tuning and testing an automated layer detection algorithm. Future adjustments in the description of an annual layer signature in the data series may also be considered. For these reasons, the inferred timescale should not (yet) be considered a final chronology, but rather an illustration of the powerful principles behind the annual layer detection algorithm developed here.

2. Visual stratigraphy of ice cores

The perhaps most basic information to obtain from an ice core is a recording of its visual stratigraphy (VS). But despite early recognition of the existence of a visible physical layering in ice cores [*Benson*, 1962; *Gow*, 1968; *Langway*, 1967], scientific use of such data proved difficult. Early studies were based on drawings of the core, and later combined with a few analog photographs [*Alley et al.*, 1997; *Meese et al.*, 1997], neither of which providing data of sufficient quality and resolution for an in-depth analysis. Furthermore, the acquired data was observer dependent, and with no opportunity to later verify the results.

With the development of relatively low-cost digital imagery equipment and increasingly large data storage media, high-resolution digital recording of the visual stratigraphy of ice cores became a possibility. The first high-resolution measurements of the visual layering of an ice core were carried out at the NGRIP ice core during the field season in 2000 [*D Dahl-Jensen et al.*, 2002; *Svensson et al.*, 2005]. Since then, similar measurements have become widely used when processing ice cores [*Faria et al.*, 2010; *McGwire et al.*, 2008b; *Takata et al.*, 2004]. Their popularity is mainly a result of the relative ease of obtaining the data, along with the measurements being non-destructive for the ice core. Yet, the extremely high level of details in such records combined with ambiguities regarding their precise interpretation, have so far limited their scientific use as paleoclimatic data series.

In this chapter, I will describe the line-scanning instrument used to record the visual stratigraphy of the NGRIP ice core, and discuss the physical origin of the visible layers seen herein. Furthermore, the chapter includes a description of how I have processed the resulting image data in order to produce a coherent gray-tone intensity curve down the ice core. Data covering the depth interval between 1866 and 2930 meters (approx. 28 to 108 ka BP) have been treated, as this depth interval contains the data of the best quality. Finally, the annual signal in the resulting data curve is discussed.

2.1 Image acquisition and quality

2.1.1 The line-scanner

At a first glance, an ice core is transparent. The ice consists of almost pure water, and only low amounts of impurities. Yet, bands of cloudy and clear ice can be observed in the glacial ice of all deep ice cores [*Alley et al.*, 1997; *Faria et al.*, 2010; *Hammer et al.*, 1978; *Svensson et al.*, 2005]. In order to obtain a high-resolution record of this faint physical layering of ice cores in much higher contrast and details than possible to see by eye, an instrument called a line-scanner was designed at the Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven, Germany, and later modified at the Niels Bohr Institute, Copenhagen, Denmark [*Nielsen*, 2005]. This instrument was carried to the field and used for recording the visual stratigraphy of e.g. the NGRIP ice core.

The line-scanner works using the principles of dark field microscopy: A dark field is placed below the ice core, and the core is illuminated by two light sources, whose beams are sent through the ice core at an angle of 45° from below (figure 2.1.1). A Charge-Coupled Device (CCD) in the scanning apparatus mounted above the core measures the amount of scattered light received. Clear ice allows most of the light beam to pass through the ice core unaffected, in which case the CCD camera records the dark field below. Areas of the ice core with a high concentration of micro-inclusions will scatter more light, and will be recorded as a bright band. This is a very efficient way to enhance the contrast of the otherwise rather subtle physical layering of an ice core.

Camera as well as light sources are mounted on trolleys. In synchrony, these move down the ice core while recording the amount of scattered light. Measurements are performed for a single line of pixels transverse to the ice core at a time, hence the name of the instrument.

In general, line-scan images from NGRIP were recorded based on 1.65 m long, 3 cm thick and 8-9 cm wide slabs of ice core. Before imaging, the surface of the ice was carefully polished on both sides with a microtome knife in order to remove the rugged surface resulting from prior cutting of the ice slab. The resolution of the line-scanner employed at NGRIP was 118 pixels per centimeter transverse to the ice core, and roughly the same along the core. The camera was an 8-bit CCD color camera. Images were labeled according to bag-number[1] of the upper part of the core section. Further description of the set-up of the line-scanning system at NGRIP can be found in *Nielsen* [2005] and *Svensson* [2005]. Examples of line-scan data from different depths in the NGRIP core are found in figure 2.2.1.

[1] The NGRIP ice core is divided into bags, with each bag having a length of 55 cm.

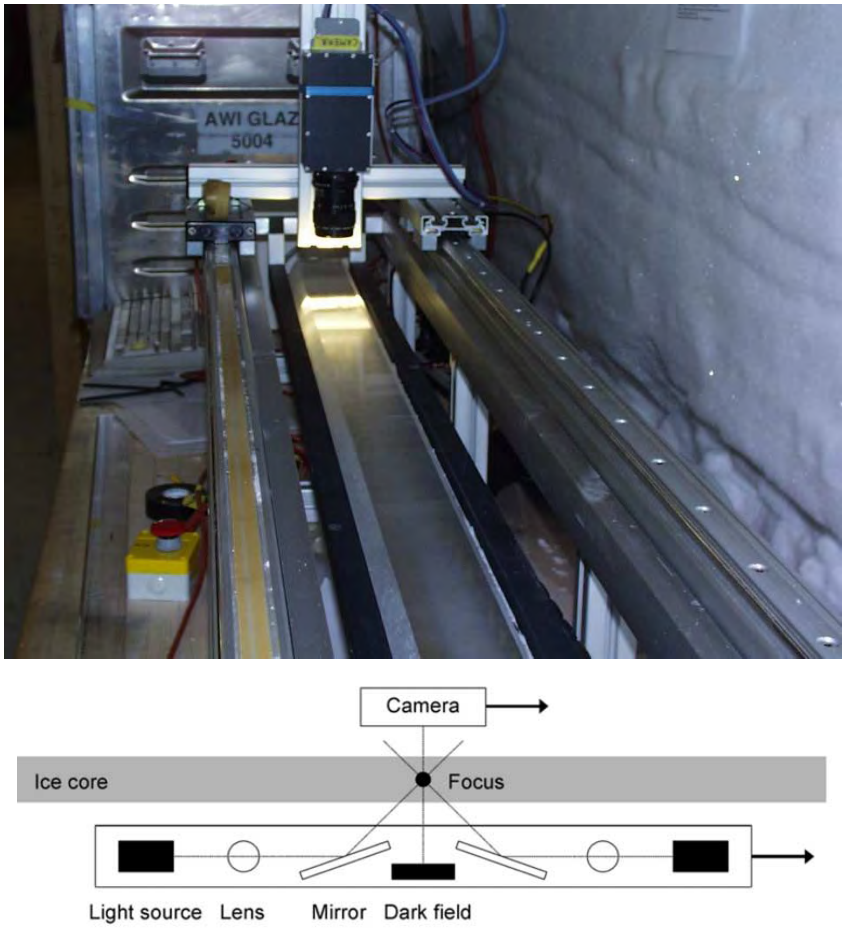


Figure 2.1.1: *The measuring principles of the line-scanner is schematically illustrated in the bottom figure: Light is being transmitted through the ice core at an angle, such that only light scattered by obstacles in the ice reaches the camera mounted above the ice core. At the top is shown the line-scanner in operation during the NGRIP field campaign. Figure reproduced from Svensson [2005].*

2.1.2 Quality of line-scan images from the NGRIP ice core

Image quality depending on storage and time

The line-scanner was first deployed at the NGRIP ice core during the 2000 field season, where the analysis was carried out on ice from the depth interval 1330-2930 m.

The upper part of the core (down to 1750 m) had been drilled during the previous field season, which only had allowed for very few scientific investigations to be carried out. For the depth interval 1330-1750 m, the core had therefore overwintered in camp for one year previous to analysis, which turned out to be a significant factor for the quality of the resulting VS data set. The extended time exposed to surface pressure had led to relaxation of the ice core, with a general evolution of air bubbles in the ice matrix originating from decomposing clathrate hydrates in originally bubble-free ice [*Pauer et al.*, 1996; *Svensson et al.*, 2005]. Scattering from these air bubbles partly obscure the overall pattern of alternating bright and dark bands in the glacial part of the VS record. The difference in image

quality from ice drilled before and during the 2000 field season is illustrated in figure 2.1.2.

Warm basal ice provided difficulties for the remaining drilling operation down to bedrock. For this reason, the visual stratigraphy of the lowest part of the NGRIP ice core (2930-3085 m) was not measured in the field but in a cold room at AWI, Bremerhaven. The measurements took place only a few months after recovery. Yet, also the VS data in these sections suffer from degradation due to the extended stay at the surface as well as temperature fluctuations during transportation.

Oversaturation of images

Due to lack of sufficient time to properly test the line-scanner before the field season, the prototype of the instrument used for recording the NGRIP VS record, turned out to suffer from a serious defect: The highest bit in the 8-bit (256 colors) CCD camera was dysfunctional. As a result, each of the three color channels restarted when reaching above the 7^{th} bit limit (128 colors).

Due to limited data storage capacities in camp (each color image took up 58MB, a gigantic file size at the time), the color images were converted to gray-scale images by taking the mean of the color channels. Apart from a few, the original images were deleted immediately after their recording.

In combination with the faulty bit in the camera, this gray-tone conversion has given rise to spurious effects in the resulting VS images. As the saturation of each of the three color channels did not happen concurrently, and each of them restarting above a saturation level of 127, the averaging process gave rise to a wide range of gray-tone intensity values being incorrectly attributed. On the original images, most of the affected areas were easily distinguishable as areas of strange coloring (figure 2.1.3A). However, such information was lost during the conversion to gray-scale images.

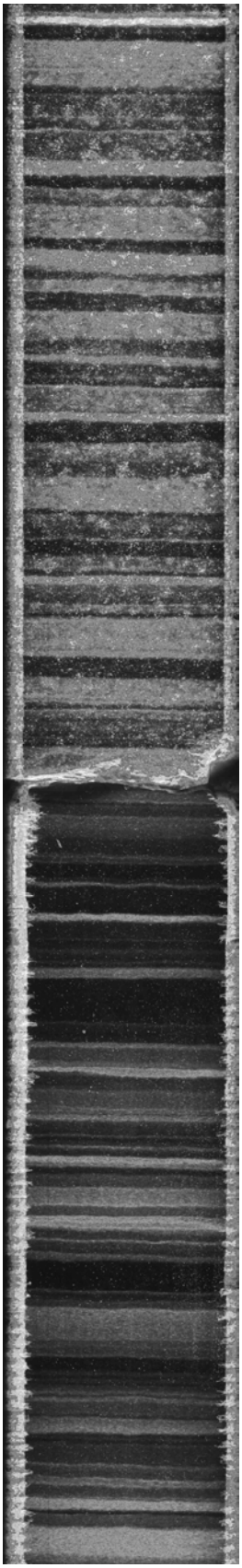


Figure 2.1.2: *55 cm of line-scan data from cold glacial ice in the NGRIP ice core (1751.20-1751.75 m), illustrating the degradation in image quality caused by storage of the ice core previous to analysis. The upper part had been stored for a year after drilling before the scan was carried out. The lower section was scanned few weeks after core recovery.*

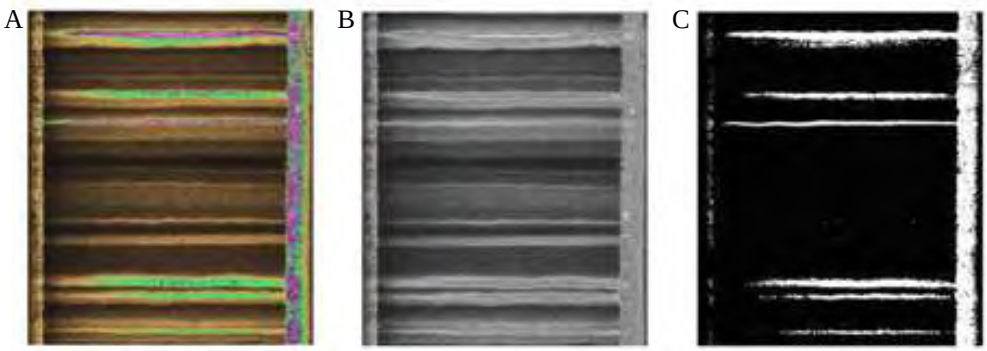


Figure 2.1.3: *A defect in the line-scan camera made it prone to oversaturation. In the original color images (A), areas of oversaturation are recognizable as being strangely colored. The gray-scale images (B) were produced by simple averaging of the three color channels. Progressive oversaturation gives rise to a lot of pixels with "intermediate grey" values. An image containing information on the location of "strangely colored pixels" (C) was produced before deleting the original color images.*

As soon as the problem was recognized, several precautions were taken in order to let this defect of the camera inflict as little as possible on the resulting data. In order to best avoid reaching saturation, the aperture of the line-scanner was repeatedly adjusted according to the changing characteristics of the ice core [*Nielsen*, 2005; *Svensson et al.*, 2005]. Despite these efforts, the VS data are for some depth intervals still severely affected by artifacts caused by saturation. Starting from a depth of 1866 m, it was furthermore recorded for which pixels in the line-scan images the coloring was abnormal, and data therefore not reliable. An example of such data is shown in figure 2.1.3C.

I will later return to the issue of oversaturation in the VS images. In section 2.3, it will be described how I have utilized the existing information to treat the VS images and have managed to recover almost flawless 8-bit gray-scale images.

Imprecise adjustment of light source

Another issue affecting the quality of the line-scan images within smaller sections was an imprecise adjustment of one of the light sources illuminating the ice core.

To fully make use of both two planar light sources, their two beams must pass through the ice core in such a way that they cross each other in the focus of the camera. In this way, they are able to supplement each other if for any reason the light from one of sources is blocked. However, one of the light sources in the line-scanner was poorly adjusted. Consequently, the uppermost couple of centimeters of every piece of core appear obscured (figure 2.1.4A), as most of the light from the active light source here was reflected by the end face of the core piece.

For similar reasons, a darkened area is often also observed around breaks in the ice core (figure 2.1.4B). The degree of shading around these regions depends on the angle of the fracture relative to the light beam.

Selected data interval

In the following, focus will be on the visual stratigraphy data from the depth interval between 1866 and 2930 meters (approx. 28 to 108 ka BP), as this is the depth interval for

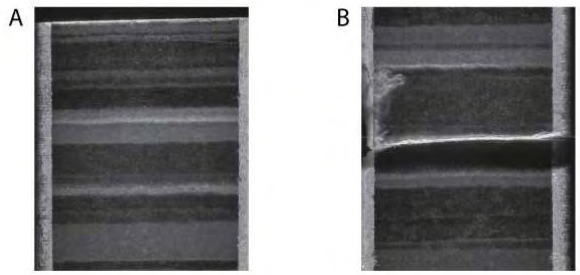


Figure 2.1.4: *Poor adjustment of one of the light sources in the line-scanner caused upper end faces (A) and breaks (B) to obscure parts of the line-scan images. Here is shown an example from a depth of 1933 m.*

which the line-scan images are of the best quality. These images were scanned under optimal conditions, namely in the field and shortly after core recovery, and furthermore care was taken to ensure that oversaturation was not too disruptive for image quality. Yet, in several sections within this interval, more than 7% of the core data are oversaturated (see also figure 2.3.1).

2.2 Visual stratigraphy as a climate record

2.2.1 What do line-scan images record?

The physical layering of the ice core as recorded in line-scan images is caused by varying amounts of microscopic impurity inclusions in the core, which are responsible for scattering the incoming light. Changes in size and/or concentration of the inclusions give rise to different amounts of scattered light [*Faria et al.*, 2010], and are seen as individual horizons in the line-scan images. The high resolution of these images makes even very thin strata (less than 1 mm) easily detectable. Most likely, each of these strata corresponds to a single deposition event.

The distinct difference between individual layers reflects the changing chemical and physical conditions on the surface of the ice sheet at time of deposition. However, the inclusions causing the scattering can either be in the form of solid impurities or air bubbles enclosed in the ice matrix, and the resulting scattering depends on their size distribution as well as quantity. Hence, the information recorded in the line-scan images is not unambiguous. Furthermore, as mentioned earlier, the signal in the core stratigraphy changes over time and according to storage conditions.

In figure 2.2.1 is found a schematic drawing of the evolution with depth of the NGRIP visual stratigraphy.

Upper part of ice core: Air bubbles

Air bubbles are very efficient scattering agents, and their variations in number and size dominate the visual stratigraphy in the uppermost bubbly part of the ice core [*Faria et al.*, 2010]. In the bubbly ice, depth-hoar sequences can be recognized based on their grain and bubble structure. Depth-hoar develops by high radiative heating of the ice sheet surface, and is hence believed to be a clear summer signal, which can be used for counting annual layers in the ice core. The Holocene part of the GISP2 time scale predominantly relies on data from visual inspection of the core, with annual markers based on a designation of

depth-hoar sequences in the ice core [*Alley et al.*, 1997; *Meese et al.*, 1997]. However, no line-scan data is available from this upper part of the NGRIP ice core.

Bubble Hydrate Transition Zone

At 900 m depth in the NGRIP ice core, the Bubble Hydrate Transition Zone (BHT) is reached. Below this depth, the steadily increasing overburden pressure causes the bubbly ice slowly to be converted into clathrate hydrates, with the air bubbles being integrated into the water molecule structure. As the refractive index of clathrate hydrates is similar to that of pure ice, this change brings about much lower scattering levels, and a distinct evolution in the visual stratigraphy is observed. At 1600 m, the last air bubbles have disappeared, hence marking the end of the BHT [*Kipfstuhl et al.*, 2001].

The evolution towards bubble-free ice does not take place uniformly. In the EPICA-DML ice core, Antarctica, it was observed that non-scattering bands devoid of air bubbles appeared in increasing number with depth in the otherwise bubbly ice. Although resembling melt layers, most of these were layers for which the transition into clathrate hydrates had taken place faster than in the adjacent ice [*Kipfstuhl et al.*, 2001]. It has been speculated that the enhanced stage of transition of these layers may be related to high impurity content, giving rise to small grain sizes and smaller air bubbles, which are disposed to a faster clathrate conversion [*Faria et al.*, 2010; *Shimada and Hondoh*, 2004]. If so, these layers – observed as dark bands in the line-scan images – may present a first step in the development towards a bright "cloudy band" as those observed deeper in the core.

Below the BHT: Cloudy bands

With the obscuring air bubbles removed, a well-defined layering consisting of dark and bright bands (so-called cloudy bands) emerges. It is widely recognized that there is a clear relation between "cloudiness" and the amount of enclosed impurities, as the cloudy bands generally correlate well with peak dust concentrations and low levels of electrical conductivity ([*Hammer et al.*, 1978; *Ram et al.*, 1995; *Svensson et al.*, 2005; *K C Taylor et al.*, 1993]). In figure 2.2.2 is shown a microstructure image of a cloudy band from the EPICA-DML ice core, Antarctica, which in details shows the increased amount of impurities in a cloudy band [*Faria et al.*, 2010].

Yet, consensus has not been reached on whether the visual banding is caused by scattering off the dust particles themselves. Some studies point to the scattering agent being a large number of microscopic air parcels (microns in diameter) which have emerged around dust grains [*Shimohara*, 2003], possibly due to relaxation of the ice core immediately after its recovery. However, other studies seem to contradict this, pointing to the actual particulates themselves being responsible: When carefully studying the melting of ice from a cloudy band under a microscope, there are no indications of small explosions formed by the release of gasses from high-pressure air bubbles [*Svensson et al.*, 2005].

In the NGRIP ice core, the end of the BHT coincides in depth with the uppermost ice from the last glacial. The distinct banding in this part of the ice core is therefore attributed to the combined effect of high dust concentration levels in glacial ice along with the disappearance of air bubbles.

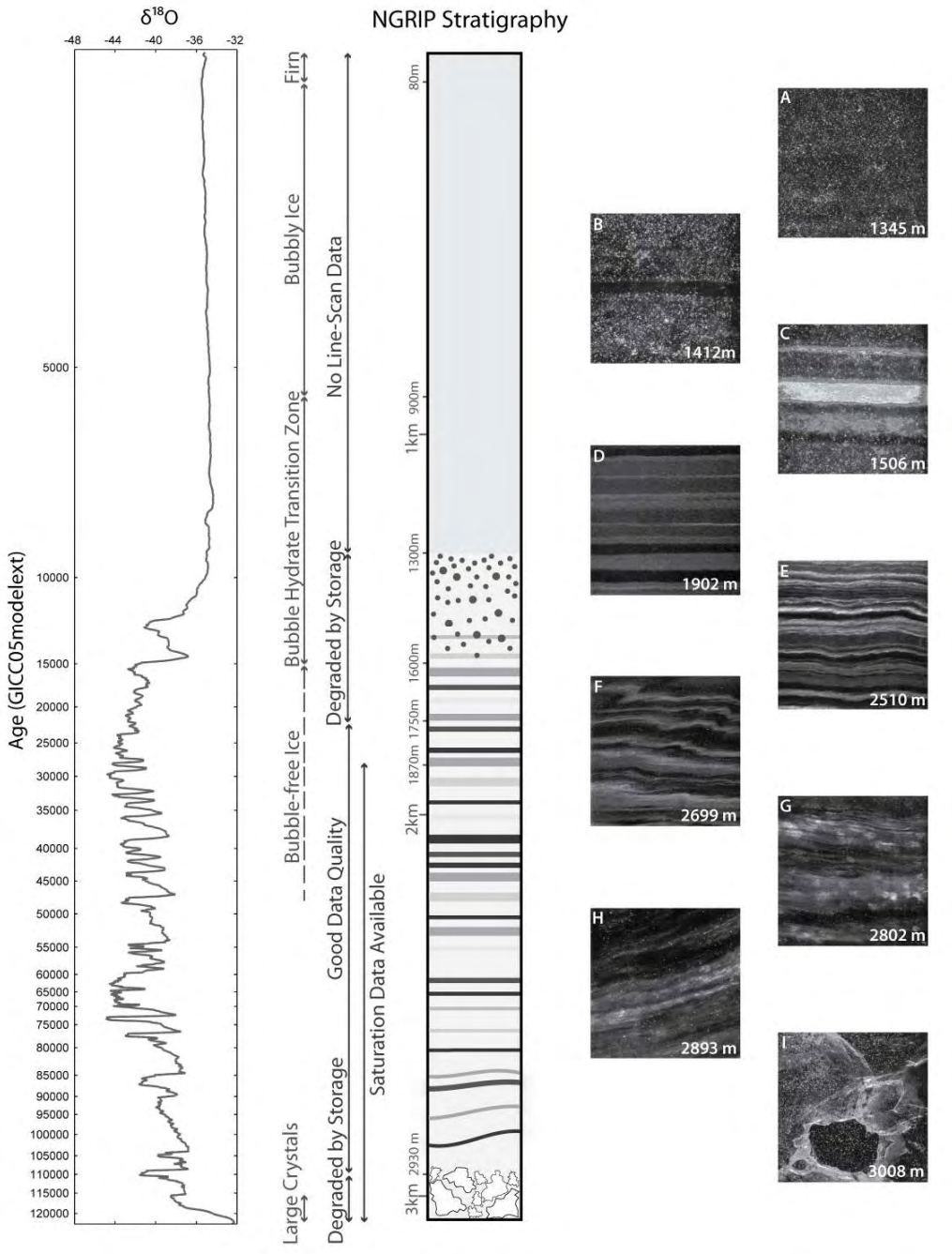


Figure 2.2.1: *Evolution with depth of the visual stratigraphy of the NGRIP ice core. In the schematic drawing, gray lines are cloudy bands, dots are air bubbles, and the bottom features are large ice crystals. Stratigraphic features are not drawn to scale. The quality of line-scan images from different sections of the ice core is marked. To the very left is shown the $\delta^{18}O$ record from the NGRIP ice core. Examples of line-scan images from different depths are shown to the right. A: Scattering from air bubbles in upper part of core. B: Sometimes a melt layer or layer of clathrate hydrates is visible as a dark band among the bright air bubbles. C: The 'Vedde' ash layer. Ice from Younger Dryas. D: Very regular horizontal banding exists for a large section of the core. E: Small-scale waviness of the dark and bright banding becomes more pronounced with depth. F: Appearance of microfolding, in places also z-folds. G: White areas resulting from crystal faces in core. H: Sections of layers with high tilt occur at great depths. I: Reflections from crystal boundaries dominate the VS in the bottom of the core. All images shown are 6.5x6.5 cm.*

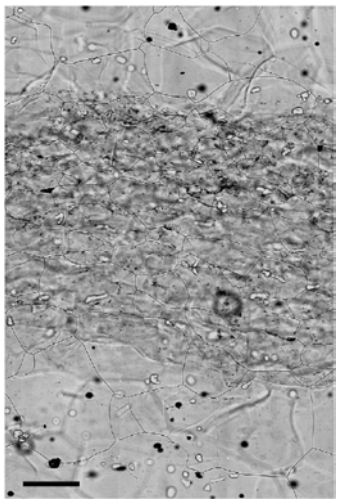


Figure 2.2.2: *A microstructure-mapping mosaic image of a 5 mm thick cloudy band from the EPICA-DML ice core, Antarctica. Depth: 1093 m. White objects are hydrates, small dark dots are micro-inclusions, and larger black areas decomposing air hydrates. Thin dark lines are crystal boundaries. Within the cloudy band, the content of impurities is larger and grain sizes are significantly reduced. The scale bar in the bottom is 1 mm. Figure reproduced from Faria et al. [2010].*

Down to great depths, the visual stratigraphy of the NGRIP ice core seems undisturbed, with nicely flat and parallel bandings. Small variations in individual layer thickness in the upper part of the core may result from sastrugi on the ice sheet surface during deposition, whereas the more evolved disturbances in the lower part of the core is due to deformation caused by ice flow. Around a depth of 2400 m, the layering starts to display a slightly wavy structure. Although continually increasing and decreasing, the degree of disturbance generally increases with depth. The first tiny z-folds appear around 2650 m. Around 2800 m, the strata in general have become rather fuzzy. At this depth, also interfaces between individual ice crystals contribute to scatter light, hereby forming slightly whitish regions in the images that disguise the underlying layering.

Bottom part: Huge crystals

In the bottom 80 m of the NGRIP ice core, high basal temperatures close to the pressure melting point have formed huge ice crystals. In this section, the line-scan images are dominated by light scattered by suitably inclined interfaces formed by crystal boundaries, and cloudy bands are no longer visible.

In the EPICA-DML ice core, similar observations were made. A more detailed analysis here revealed that it was not just that the scattering from crystal boundaries was obscuring an underlying dark and bright banding pattern. The micro-inclusions previously forming the cloudy bands was aggregating at grain boundaries and hydrates, gradually leaving the ice in between cleaner and cleaner [*Faria et al.*, 2010]. Such impurity migration may have high impact on the details of the paleoclimatic data series from the ice core. However, it seems that a similar redistribution of impurities does not take place at NGRIP. Here, high-resolution impurity measurements revealed the existence of annual layering even in the very bottom part of the ice core [*Svensson et al.*, Submitted 2011].

2.2.2 Information on layer disturbance

Besides from information contained in the varying brightness of the banding pattern, the visual stratigraphy also reveals the variations in layering across the core diameter.

For most ice core measurements (this is e.g. the case for CFA data), a single value is assigned to each depth, hereby assuming such variations to be negligible. The reported value constitutes an average of a cross-section of the core, with an averaging width depending on the specific instrument. However, as mentioned earlier, although the banding pattern is consistently flat and regular for the upper part of the NGRIP ice core, waviness and z-folds of individual layers is a reoccurring phenomenon in the lower 500 m of the core. In this case, a reported average value across the core may have smeared out the variability in the high-resolution signal of the data set.

In this context, however, it should be emphasized that also the visual stratigraphy in the line-scan images constitutes an averaging. After all, the images are only two-dimensional, and during scanning, the line-scanner focuses over a small depth interval within the prepared ice core slab. As a consequence of this averaging, the layer boundaries gradually becoming less distinct with depth and increased waviness of the general layering structure.

It should also be noted that the observed folding and waviness of the layering depends on the surface of core which was prepared for analysis. During extraction of an ice core, information on its absolute orientation in the bore hole is not preserved, and the surface to be prepared for scanning is selected more or less arbitrarily. Hence, the disturbance in the visible layering in consecutive sections of the core may not be similar, each of them providing a lower limit on the degree of disturbance only.

Information contained in the degree of small-scale disturbance of the layering may indeed have more far-reaching implications than simply a decrease in resolution of the individual ice core records. A substantial degree of small-scale disturbances may signify disturbances on much larger scales, and may warn about disruptions in the stratigraphic continuity of the paleoclimatic records. The divergence in the lower part of the climate records from the GRIP and GISP2 ice cores coincide with an increase in small-scale disturbance of the layering in both cores, and similarly was observed for the EPICA-DML ice core in Antarctica [*Faria et al.*, 2010]. Despite the general increase in layer inclination and waviness with depth of the NGRIP ice core, however, the NGRIP record seems to be undisturbed all the way to the bedrock [*North Greenland Ice Core Project Members*, 2004].

2.3 Processing of line-scan images

To facilitate the subsequent analysis, a gray-tone intensity profile was extracted from the line-scan images. However, to retain as much as possible of the highly detailed data record contained in these images, several issues must be kept in mind when constructing such intensity curves. As discussed in *McGwire et al* [2008b] and *Katsuta et al* [2003], averaging image intensities over several pixels reduce the noise level of the resulting intensity profile. But ideally, in order to preserve the high data resolution, only intensities belonging to the exact same horizon should be averaged.

Another key consideration concerns the relative calibration between intensity profiles obtained from line-scan images recorded with different apertures. For the performance of the automated annual layer counting routine (chapter 3), it is crucial that the character of the data series does not change too abruptly with depth – and indeed; not as a result of technicalities in the measurement procedure.

Additionally, the NGRIP line-scan images must be corrected for artifacts caused by the malfunctioning CCD in the line-scan camera. Data quality may otherwise be significantly degraded by oversaturation, the degree of which depending on data characteristics as well as the employed aperture value (figure 2.3.1A).

Other data treatment includes dealing with breaks in the ice core, and the construction of a depth scale for each core section scanned.

2.3.1 Selecting image data areas

Locating edges of ice core

First of all, we need to be able to distinguish the actual ice core on the line-scan images from the background. For this purpose, an image-processing tool was developed which automatically could locate the core boundaries.

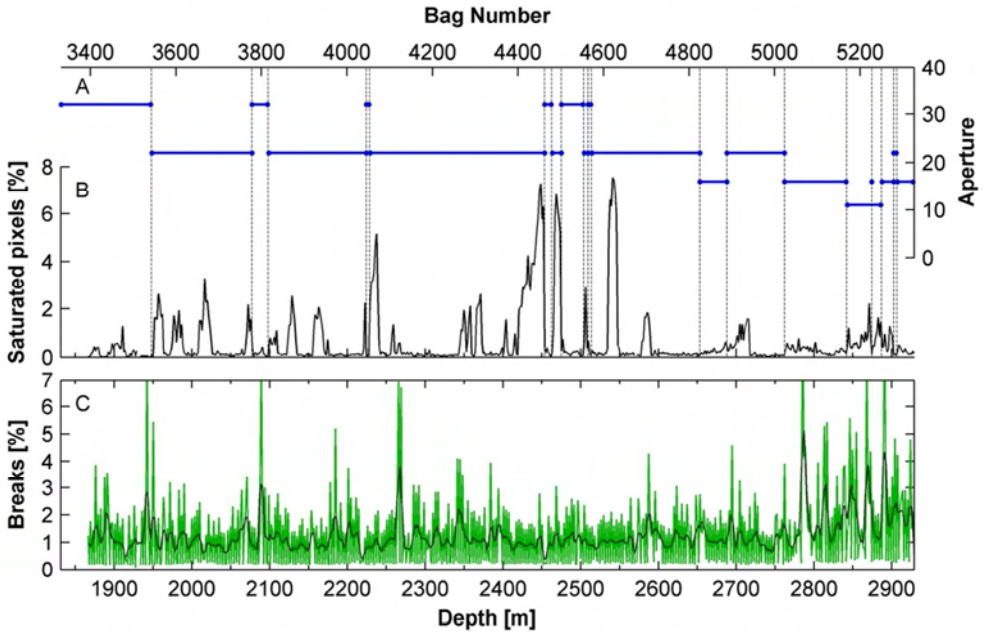


Figure 2.3.1*: A) and B): Applied aperture values for line-scan images within the selected depth interval, and the resulting percentage of saturated pixels observed. Only areas occupied by ice core in the images are considered. The percentage of saturated pixels provides a lower bound only, as pixels experienced full saturation in all color channels are not included. Not surprisingly, a change in aperture is most often associated with a sharp decrease/increase in saturation. C) Percentage of line-scan image data which had to be masked out due to breaks in the ice core, overexposure etc. When looking in details, a regular zigzag pattern emerges. This is due to stable drilling conditions with very regular lengths (up to 3.55 m [D Dahl-Jensen et al., 2002]) of ice core sections retrieved at each run. As this is close to twice the length of the ice core sections scanned at a time, approximately every second image contains a break.*

Although not always explicitly acknowledged in the following, knowing the location of core boundaries plays a significant role in many of the subsequent image processing steps. Together with knowledge on the precise length of the individual ice core sections scanned, it is also utilized for constructing a depth scale for the visual stratigraphy data.

Alignment of core sections

In the considered depth interval, the layering in the NGRIP ice core is more or less horizontal. However, as the ice core itself often not is placed straight in the core-scanning device, layers are not always horizontal in the image itself. Given that intensity profiles are constructed by horizontal averaging of intensities, proper alignment of the core in the line-scan images increases the resolution of the resulting data series. Besides, the aligned and cropped images are in general much handier to work with than the original data.

The images were aligned and cropped in the following way: For each unbroken piece of ice core, a straight line was fitted to the vertical edges of the core, and the core piece was subsequently aligned using the corresponding angle. Afterwards, parts of image only containing background data were removed.

Potentially, the aligned core data could further be used for describing how the layers in the core changes with depth.

Locating breaks in core

Another image processing tool was developed to locate breaks in the core data. Such breaks can be recognized as areas of the core, which are bright and with a different inclination to horizontal than the surrounding layers. Furthermore, due to the one inactive light source in the line-scan set-up, breaks in the ice core leave a very specific signature in the core edges, in which a dark area is imprinted.

Based on the above observations, a filter has been created, which works on the aligned data to create a “bad ice mask” containing regions of the ice core which are breaks. However, as breaks can be very hard to discern from bright layers with a slightly wavy structure, the constructed mask was inspected manually afterwards to ensure that the correct areas had been masked out. The “bad ice mask” both contains the actual break in the core and potential dark areas below, for which the break blocked out most of the light from the active light source.

The very top and bottom (~1 mm) part of the core did not contain useful data either, as the areas here were oversaturated due to increased scattering from the edges. Such areas were easy to identify due to their high intensity values, and an algorithm was developed to mask out data also from these regions.

On average, a little more than 1% of the image data had to be removed due to breaks and scattering off the ice core end faces.

2.3.2 Reconstructing 8-bit line-scan images

Comparing raw gray-tone intensity data

A few line-scan images within the considered depth range were taken more than once and using different apertures. A compilation of these and the applied apertures is found in

Image	Aperture			
	f/11	*f*/16	*f*/22	*f*/32
3778			×	×
4045			×	×
4462			×	×
4501			×	×
4564			×	×
4825		×	×	
4888		×	×	
5023		×	×	
5227	×	×		
5278		×	×	

Table 2.3.1: *Line-scan images captured using more than one value of aperture. Within the considered depth interval, 10 such images exist.*

table 2.3.1. From a comparison between such images, the effect of the dysfunctional bit in the CCD camera combined with the applied gray-tone conversion of the images can be observed.

In figure 2.3.2 is plotted the relationship between the raw gray-tone intensities of the five line-scan images measured both with apertures of *f*/22 and *f*/32. Obviously, the relationship between the two sets of intensities is non-trivial. An aperture of *f*/32 is the smaller one, resulting in darker images and therefore a lesser degree of saturation. Consequently, they provide the better estimate of the actual intensities. An aperture of *f*/22 is twice as large as that of *f*/32, allowing twice as much light to enter the CCD, and hence the theoretical relationship between intensity values measured with those two apertures is a straight line with a slope of approximately 0.5. This is also observed for low intensity values. However, for intensities above 60, the linear relationship breaks down. Observe how much of the dynamical range is mapped into a region of intensity values around 60 in both images, thereby giving rise to a large amount of ‘intermediate gray’ colors in the line-scan images.

In figure 2.3.3 is compared a small section of two intensity profiles from the same depth interval, but based on line-scan images obtained using different aperture values. The one captured with an aperture of *f*/32 has experienced very little saturation. Oversaturation of the image scanned with larger aperture (*f*/22) changes the peak values in the gray-tone intensity profile. Caused by the non-linearity of the relationship between original and observed intensity values, peaks may even be turned into valleys, thereby significantly degrading the overall signal in the data series.

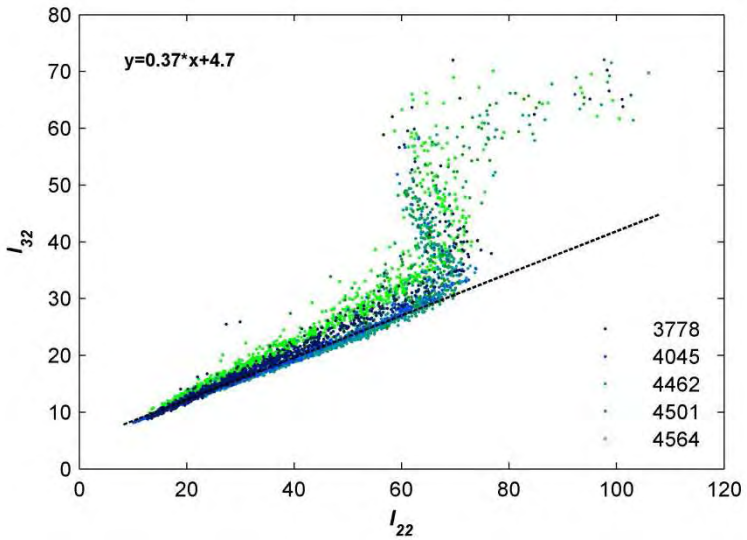


Figure 2.3.2: *A comparison of pixel-to-pixel gray-tone intensities for the line-scan images scanned both with aperture f/22 and f/32. For intensities below 60, a linear relationship is found. The best fitting straight line for these values is plotted. Above this value, however, the linear relationship breaks down, and intensity values are clustered in the interval between 60 and 70. The development of a secondary linear branch for high intensity values is caused by full saturation of the large-aperture image (see discussion in text). For the five images selected here, 30% of all intensity values in the large-aperture images are above 60 and hence possibly unreliable. For reasons of discernibility, only a small subset of the available data is shown. To account for small discrepancies in the co-registration, data has first been smoothed with a Gaussian filter of width 5 pixels.*

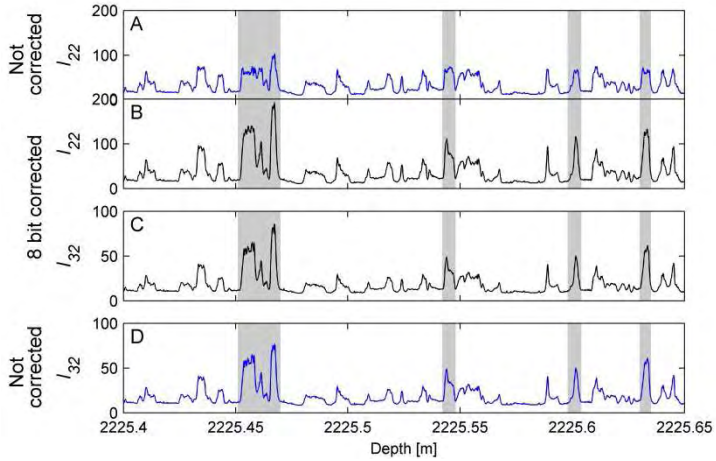


Figure 2.3.3: *Corrected and uncorrected intensity profiles. The intensity profile measured with aperture f/22 (A) is much more affected by oversaturation than that measured with an aperture of f/32 (D). Note in A, how all peak values are more or less similar. As seen from B (or D), this is not a true phenomenon. In general, oversaturation causes the peaks to be less pronounced. The gray bands in particular are heavily influenced. Also note how e.g. the very top of the peak around 2225.65m has turned into a small depression by the combination of oversaturation and color channel averaging.*

Correction of gray-tone intensities

Having realized the importance of saturation for the visual stratigraphy data quality and the coherence between subsequent line-scan images, the unsaturated 8-bit line-scan images have been reconstructed on a pixel-to-pixel basis.

The effect of the bad bit on the recorded and subsequently averaged line-scan images can be realized from the following line of reasoning: For each of the three color channels (here denoted by r, g, and b) the camera went into saturation for values above 127, at which point the color scale restarted at 0. That is, with $X_i \in \{0, 1\}$ for $i \in \{r, g, b\}$, the observed averaged intensity in a specific pixel is related to the original intensity value by:

$$\begin{aligned} I_{obs} &= \frac{1}{3}\Big((r - 128 \cdot X_r) + \big(g - 128 \cdot X_g\big) + (b - 128 \cdot X_b)\Big) \\ &= \frac{1}{3}(r + g + b) - \frac{128}{3} \cdot \big(X_r + X_g + X_b\big) \\ &= I_{true} - \frac{127}{3}\big(X_r + X_g + X_b\big) \end{aligned}$$

Consequently, the intensity value to be observed, had the camera not been malfunctioning, can be calculated as:

$$I_{true} = I_{obs} + \frac{128}{3}\big(X_r + X_g + X_b\big) = I_{obs} + \frac{128}{3}X^* \qquad (2.3.1)$$

Although we do not know the individual correction factors X_i corresponding to a given pixel, the total correction (X^*) to apply to the observed intensity values can only be one out of four possible values:

$$X^* = \big(X_r + X_g + X_b\big) \in \{0, 1, 2, 3\}$$

With increasingly high intensities, the image gets progressively saturated, and a larger and larger saturation factor is to be applied. The combined effect of increased intensities and increased correction factors is that for a large range of values of original intensity values, more or less the same intensity is observed. Finally, having reached full saturation, the correction factor can no longer increase, and the observed intensity values will again increase. Hence, for sufficiently high intensities, we would in figure 2.3.2 have observed a straight line with the same slope as the one observed for low intensities, had not the small-aperture image also reached saturation at these values.

To obtain a guess of which correction factor to apply to which pixel, the recorded information on "strangely colored pixels" (see example in figure 2.1.3) was used together with an assumption of smoothly varying intensity values.

Most pixels in the line-scan images need no correction. Yet, areas of strange green or purple coloring are caused by respectively one or two color channels being over-saturated[2]

[2] Unfortunately, it was not recorded which pixels were in green colors and which were purple. Such information would have been a great help for correcting the images. The conversion between colors and correction factors appears to be straight-forward, with green colors corresponding to a correction factor of 1, and purple corresponding to a correction factor of 2 (see figure 2.1.3A).

Consequently, the appropriate correction factor here is limited to being either 1 or 2. If all three color channels are oversaturated (hence requiring a correction factor of 3), the colors are again in balance, and the result is grayish like the original data. Assuming slowly varying intensities in all color channels, the correction factors are likely to vary one step at a time. Hence, the only pixels which may require a correction factor of 3 are located within an area surrounding by saturated pixels. However, it may also be that such regions are not saturated at all. Thus, possible correction factors for these areas are 0 and 3. The rest of the image is left uncorrected.

Subsequently, correction factors were assessed for those parts of the line-scan image hereby identified as potentially oversaturated. By starting from the border of each such area and slowly filling it in, the most likely correction factor for each pixel was found. The likelihoods of both possible correction factors were assessed based on the similarity between the resulting gray-tone intensity and a weighted mean of known intensity values in the immediate surroundings. As the intensities generally are more alike parallel than perpendicular to the layering, largest emphasis was placed upon the similarity of intensity values in the horizontal direction. Pixels with assigned correction values were hereafter treated as having known intensities, and were used for estimating the remaining corrections.

As the above described reconstruction procedure continuously uses information from correction factors previously obtained, a decrease in quality of the reconstruction may be expected with increasing size of area to be filled out. To circumvent this, the estimated best correction values were accepted only where these were unambiguously defined. Pixels with no significant difference in likelihood of the two possible correction values were not assigned any of these. They had to wait for the correction values of other less ambiguous pixels in the neighborhood to be determined. With increasing number of surrounding intensity values known, a better estimate of the most likely correction factor for the pixel in consideration could be made. In this way, uncertain corrections were not allowed to inflict significantly on the resulting reconstructed image. The level of certainty required for assigning a correction factor to a given pixel was successively lowered as the algorithm started struggling to find correction factors good enough to be accepted.

To further increase the robustness of the procedure, the algorithm was run twice on each image. After having obtained a first estimate of correction values, regions of non-smooth intensities changes were mapped. The algorithm was then re-run a second time in a neighborhood around these areas.

Performance of reconstruction procedure

In figure 2.3.4 (and figure 2.3.6) is shown the relationship between corrected intensity values of images captured using more than one aperture value. As desired, the relationship is now linear for the entire range of intensity values. The improvement relative to the same relationship for uncorrected images (figure 2.3.2) is evident.

Given that an aperture of *f*/22 is twice as large as *f*/32, the theoretical slope of the two sets of intensity values is 0.5. However, these aperture values are likely to be taken more as an estimate of the actual aperture area rather than an exact number. From the observed slopes, it seems that the aperture setting on the line-scanner was not precisely 0.5, but

rather 0.4. Furthermore, as the aperture was adjusted manually, the value was not exactly the same each time, which can be seen from the slightly different slopes of the best fitting straight lines when based on the individual line-scan images.

With the above procedure, the line-scan images could be reconstructed almost perfectly in full 8-bit resolution. If looking carefully at each line corresponding to a specific image, it seems that for very high intensity values, the algorithm may tend to assign too small correction values to a minor fraction of the pixels, hereby giving rise to a slight underestimation of the height of very high peaks. However, the differences are generally insignificant.

Visual inspection of the line-scan images revealed that while thin layers having experienced oversaturation are truly impeccably restored, few imperfections may exist for very thick layers, as these required a lot of processing in order to be restored. High intensity values most often correspond to such thick layers, and this may be the reason for the slight underestimation of very high intensity values mentioned above. In figure 2.3.5 is shown a comparison between the original oversaturated image and its corrected version.

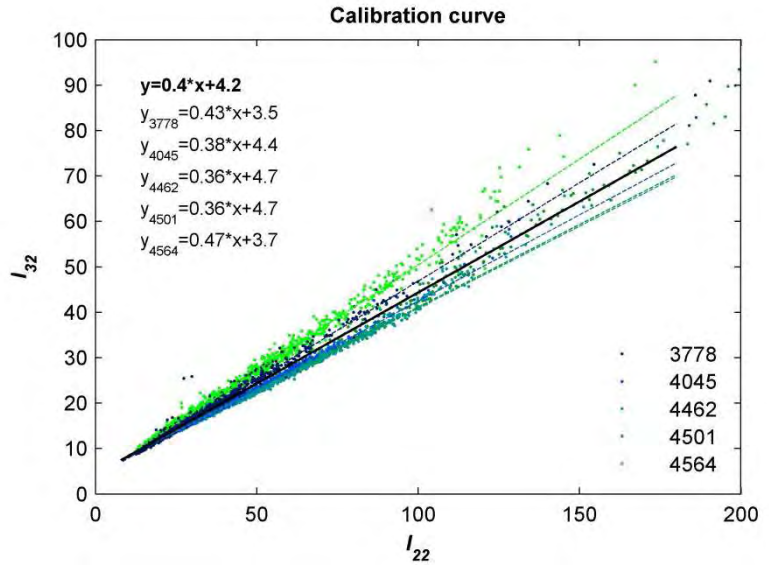


Figure 2.3.4: *Corrected intensity values for the same line-scan images as shown in figure 2.3.2, namely those measured both with aperture f/22 and f/32. After the correction procedure, the relationship is a straight line for all intensity values. Only a small subset of the available data is shown, whereas the entire data set has been used to calculate the best fitting straight lines. Most of the scatter stems from difficulties co-registering the data on a pixel basis. To account for small discrepancies in the co-registration, data has first been smoothed with a Gaussian filter of width 5 pixels.*

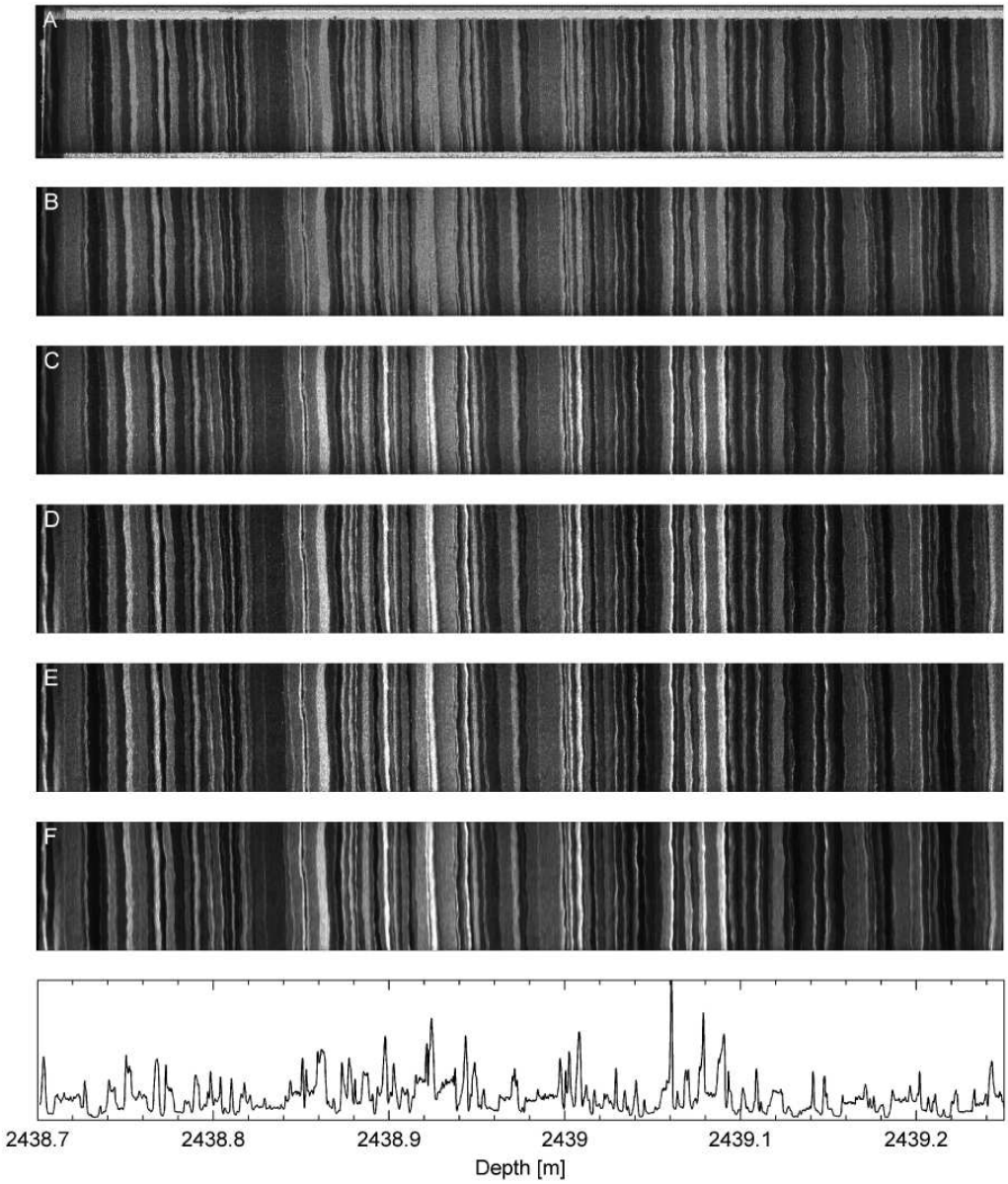


Figure 2.3.5: *The image processing steps. The original line-scan images (A) are cropped and aligned (B), corrected for the dead pixel and the faulty 8th bit (C) and corrected for illumination effects (D). Finally, white speckles are removed (E) and the image is median filtered to remove noise (F). The resulting intensity profile is found in G.*

2.3.3 Other image improvements

Correcting for dark current

A Charge-Coupled Device (CCD) detects the received amount of light by capturing photons, turning them into electrons, and counting the accumulated number of electrons in tiny bins. Even when left in complete darkness, however, CCDs will collect some electrons. Hence, on the read-out of image, a so-called dark current will be present, providing a minimum level of measured intensities.

The dark current is independent of aperture. Hence, the dark current level for the employed CCD can be found from the relationship between measured intensity values using different apertures (like figure 2.3.4 above). Using again apertures *f*/22 and *f*/32 as an example, we have:

$$I_{22}^{corr} = I_{22}^{obs} - I_{dark}$$

A similarly relationship holds for I_{32}, leading to:

$$I_{32}^{obs} - I_{dark} = I_{32}^{corr} = kI_{22}^{corr} = k\left(I_{obs}^{22} - I_{dark}\right) \Leftrightarrow$$

$$I_{32}^{obs} = kI_{obs}^{22} + (1-k)I_{dark} \tag{2.3.2}$$

From the best fitting straight lines, the dark current level of the line-scan images can therefore be determined. Its value is found to be close to 7, and this value has therefore been subtracted from the observed intensity values.

Varying lightening conditions

Having first subtracted the dark current, we can correct for different illumination conditions across the image (flat-field correction). By itself, this correction has limited influence on the shape of peaks and troughs in the resulting intensity profiles, as it mainly results in a general change of level of intensities measured. However, for the "bad ice mask" (see p. 22) to work properly, it was necessary first to apply such a flat-field correction to the images. It also increased the performance of the reconstruction algorithm, which assumed slowly varying horizontal intensity changes.

Looking at the line-scan images in details, it is clear that ice core was not uniformly illuminated during scanning. Individual layers do not maintain the same color across the core, but appear darker towards the edges. As the trend seems to be more related to position in the image than to core edges, the effect seems mainly to be caused by the light sources used to illuminate the ice core from below rather than shadowing of light by the core edges.

The general trend in gray-tone intensities across the image was found by vertically averaging the measured intensities in the line-scan images. Repeating this for many images and averaging their normalized trends, the resulting curve gave an estimate of the general pattern of intensity values across the ice core. The effect of changing lightening conditions could then be removed by dividing the intensities with the appropriate correction factor (figure 2.3.5D).

A similar exercise was done for the upper part of the line-scan image from each ice core section, which suffered from decreased illumination due to high amounts of scattering of the active light source from the nearby core edge. In this case, however, the decrease in intensity does not depend on the position in image, but the distance from the edge of the core. Also, the pattern is not constant across the core. Nevertheless, from the averaging of many co-registered images, a general illumination pattern was obtained and its effect could be removed (figure 2.3.5D).

Dead pixel

The CCD camera used in the line-scanner at NGRIP had an inactive pixel, which can be recognized as a dark vertical line on each line-scan image. The appropriate intensity values here have been interpolated as the average intensity of the pixels on either side.

Noise filtering

The line-scan images are generally filled with white speckles, most of which probably artifacts produced from polishing the ice core with a microtome knife [*Faria et al.*, 2010]. These tiny bright spots are easily recognizable on the line-scan image and have been removed. Their new intensity values were found as the average intensity of the surrounding pixels (figure 2.3.5E).

Finally, the image was filtered using a median filter in order to remove as much noise in the image data as possible before constructing the intensity profile (figure 2.3.5F and G).

2.3.4 Constructing gray-tone intensity profiles

Obtaining intensity profiles

Finally, the intensity profiles were constructed based on the aligned images, and measured down the center of the core. As the visible strata in the NGRIP line-scan data are more or less horizontal throughout the considered depth interval, the intensity profiles were constructed as a simple mean of 50 intensity values perpendicular to the ice core.

No efforts were made to account for the increased tilting and waviness of the individual layers with depth[3]. Breaks and other areas without useful data were disregarded when calculating the averages.

Aperture calibrations

All intensity profiles were then calibrated to the one obtained with an aperture of *f*/22. For the depth interval under consideration, this is the aperture most commonly used (figure 2.3.1). The calibration factor (‘*k*’ in eq. (2.3.2)) corresponding to different aperture values were found from the comparison between two exposure-corrected scans of the same ice core (figure 2.3.4, figure 2.3.6). The calibration factors are found as the slope of the best fitting straight line when using all data points from all images available.

By applying the aperture calibration to the resulting intensity profiles, and not directly on the images themselves, we allow for non-integer data values, as well as intensity values above 255.

[3] A detailed method of doing so has been outlined in e.g. Katsuta, N., M. Takano, T. Okaniwa, and M. Kumazawa (2003), Image processing to extract sequential profiles with high spatial resolution from the 2D map of deformed laminated patterns, *Comput Geosci-Uk*, *29*(6), 725-740.

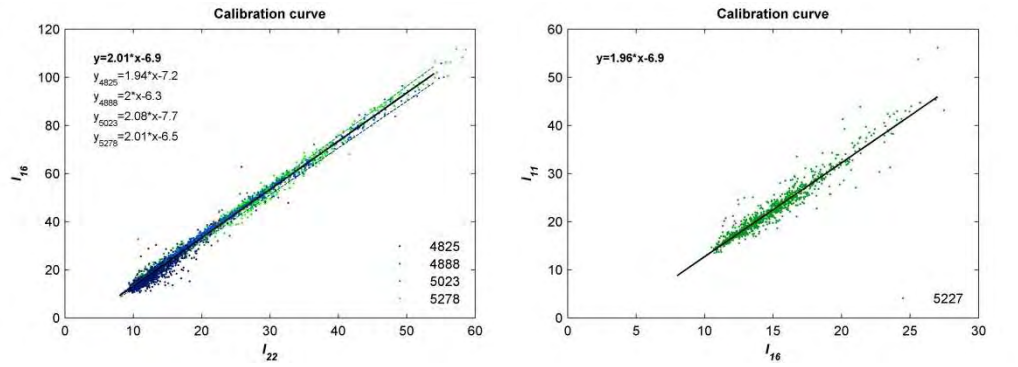


Figure 2.3.6: *The linear relationship used for calibrating between I_{16} and I_{22} (A), and from I_{11} to I_{16} (B). Intensity profiles measured with an aperture of f/11 can be calibrated to an aperture of f/22 by first applying the calibration factor relative to I_{16}, and then the one from I_{16} to I_{22}.*

2.3.5 Depth scale of line-scan images

Line-scan images directly provide a "proper" depth scale. Disregarding breaks in the ice core, which always present a source of depth uncertainty, the accuracy of the line-scan depth scale is only limited by the steadiness of movement of camera trolley in the line-scanner. Assuming a constant velocity, the depth scale can be established based on the length of the ice core section in the image, and its true length. The involved uncertainties in the constructed depth scale are difficult to estimate as they entirely depend on the core section in question. For most core sections, however, these are negligible.

In the following, we will often wish to compare the visual stratigraphy with data from e.g. CFA measurements. However, a detailed comparison of such two high-resolution records is severely hampered by small differences in the two depth scales. Whereas the visual stratigraphy associates an absolute length scale to the core, the depth scale derived from the CFA system is slightly harmonica-like. It depends on the melt rate of the ice rod, which is difficult to control and which may slightly change during the measurements, as well as on the travel time of the water stream through the pump, tubes and analytical lines. The resulting depth scale may be up to several centimeters off – a large difference when comparing to annual layers with a thickness of the same order of magnitude. The depth scale variations are even larger with ECM data, which are measured by hand.

2.4 Annual layer signal in VS

For the considered depth interval of the NGRIP ice core, the visual stratigraphy displays a clear banding of dark and bright layers, and the derived intensity profile resembles the dust concentration signal in great details [*E. Kettner, pers. comm.*]. Due to the seasonality of inclusion of dust into the ice [*Alley et al.*, 1997; *Hamilton and Langway*, 1968], cloudy bands in the visual stratigraphy have the potential to be used as indicators for annual layers.

The present seasonality of dust concentration in deposited snow is believed to be the result of intensified dust storms occurring during spring and summer, in combination with seasonal changes in atmospheric circulation [*Ram and Illing*, 1994]. But increased dust

concentrations may occur throughout the entire year due to e.g. irregular changes in weather patterns or influx of volcanic dust.

Similar to the irregular seasonality pattern expressed in dust profiles, the annual signal in the visual stratigraphy is not very regular. In Holocene ice, it has been observed that often more than one strong cloudy band is formed during springtime, and that a secondary layer of increased dust content often occurs in fall [*Alley et al.*, 1997]. From considering the visual stratigraphy data from NGRIP, its seasonal pattern generally seems to be more regular during the cold periods than during the warm interstadials.

In figure 2.4.1 is shown the result of a spectral analysis on visual stratigraphy data from the Greenland Stadial 13 (GS-13). The seasonal pattern is too irregular to show up as a peak if considering the data on a depth scale (figure 2.4.1A). However, a seasonal signal does exist in the data. This can be seen from a spectral analysis on the data on a timescale according to GICC05 (figure 2.4.1B).

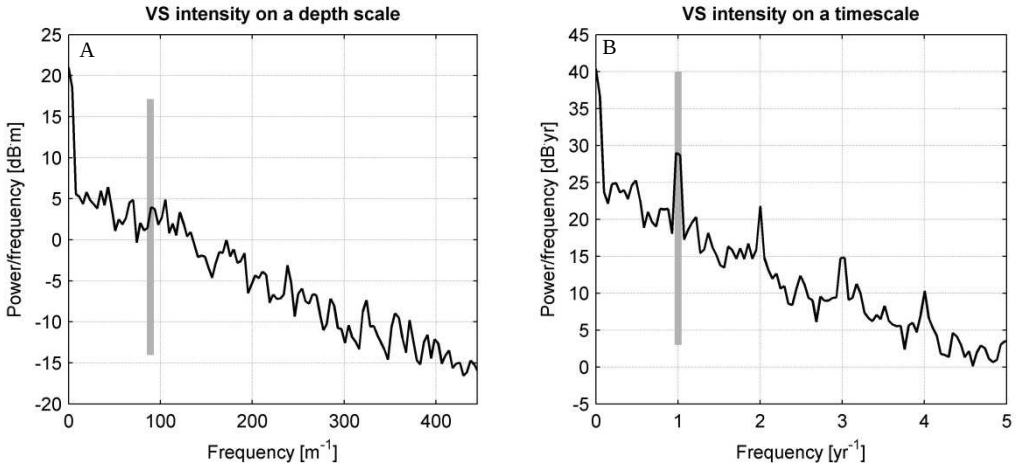


Figure 2.4.1: *Spectral analysis on the corrected visual stratigraphy profile from a depth of 2233 m in the NGRIP ice core (GS-13). A: Spectral analysis on the data on a depth scale. B: Spectral analysis on data on a timescale according to GICC05. The gray bar signifies the location of an expected annual peak. In this spectrum, several other regular peaks with lesser importance also show up. These may be attributed a non-cosine structure of the annual layer signal, causing it to require more than just one spectral component to describe the signal.*

The most extensive use of visible strata for deriving an ice core time scale was for the dating of the GISP2 ice core. The Holocene part of the GISP2 time scale predominantly relies on data from visual inspection of the ice core, with annual markers based on a designation of depth-hoar sequences in the core. Further down the ice core, where depth-hoar sequences were no-longer visible, visible cloudy bands along with other annual layer indicators (ECM, laser-light scattering from dust) were used for establishing an annually counted timescale for the glacial part of the ice core. Cloudy bands in the visual stratigraphy were then used to extend the dating of the GISP2 ice core back to 50 ka BP [*Meese et al.*, 1997]. To a lesser extent, the VS profile was also used as one of several parameters for the deeper part of the GICC05 timescale.

3. Layer counting using Hidden Markov Modeling

Establishing a chronology for a paleoclimatic record by manual counting of annual laminations is a tedious and furthermore subjective task. The Greenland Ice Core Chronology 2005 (GICC05) was developed over several years, involving the persevered efforts of many researchers; counting, comparing and re-counting the layers in the ice core data. To reduce the subjectivity of layer counting, and hence increase the quality of the resulting timescales, many attempts have been made to develop automated methods of doing so. However, this is not an easy job, and most methods have since been abandoned, leaving the manual layer counting approach to still be the most accurate.

The attempts range from simple approaches mainly concerned with smoothing the data beforehand and simple counting of the remaining peaks [*Shimohara*, 2003] to those more elaborate and sophisticated [*Rasmussen et al.*, 2002]. General features, however, are preprocessing of the data in form of smoothing or bandpass-filtering (hereby implicitly using some prior knowledge on the involved layer thicknesses), in combination with either a one-layer-at-a-time approach [*McGwire et al.*, 2008a; *Rasmussen et al.*, 2002; *Smith et al.*, 2009] or a general search for periodicities in the data [*Rupf and Radons*, 2004; *Svensson et al.*, 2005]. In comparison to these, the method developed here can be regarded as a one-section-at-a-time approach, where multiple layer boundaries in an entire data section are being determined simultaneously. This is an approach much more similar to the manual approach of layer counting.

The method developed here is a novel method of automating the annual layer counting procedure. It is based on the statistical framework of Hidden Markov Modeling (HMM), which originates from speech recognition, and to my knowledge has not yet been applied in any paleoclimatic context. In many ways, the layer detection methodology described here resembles what is automatically done by eye. And although the human eye is almost unsurpassable when it comes to pattern matching, an automated approach will always benefit from its objectivity.

Before knowing about the existence of the Hidden Markov Model framework, I tried out various other methods. However, too large variations in the individual annual layer thick-

nesses resulted in poor performance for methods looking for specific frequencies in the data. And methodologies based on a one-layer-at-a-time approach simply lacked the ability of using the entire data sequence to best pick out the layer boundaries.

Given the large variability in expression of the annual layers and the high degree of noise, it is often not possible to detect all layers with certainty. When using a one-layer-at-a-time approach, the accumulation of errors from wrongly positioned annual layer boundaries turned out to be crucial for their performance, and I could not make such algorithms work properly. In contrast, the present algorithm is able to position the annual layer boundaries based on the entire data section at once. The layering in sections with poorly resolved annual layer peaks is therefore determined based on the positioning of clearly discernible layers before as well as after. In this way, a false peak will not disrupt the overall performance of the layer counting routine, hereby making it much more robust against noise and variability of the annual layer signal of the data in question.

This chapter will start out with a short introduction to Bayesian statistics, which provide the fundamental theoretical basis for the approach. The layer detection routine itself is based on the Bayesian statistical framework of Hidden Markov Modeling (HMM). The concepts of this framework will be presented, and it will be described how they can be applied to the case of annual layer detection in ice cores. Assuming the observations to be the outcome of a hidden Markov process, layer detection in an observation sequence can then be achieved by using one of two algorithms: The Forward-Backward and the Viterbi algorithm. Although very similar, they are based on different perceptions of what is the 'best' annual layering in a given data sequence, and will therefore give slightly different results having slightly different interpretations. Equations related to annual layer detection will be derived for both algorithms, and their differences will be discussed. The layer detection routine is developed with the application for visual stratigraphy in ice cores in mind, but the general concepts can be used for a wide range of similar applications.

The layer detection algorithm works by reducing the complex issue of simultaneous pattern matching of multiple successive layers to a given template, to the much simpler question concerning how likely a particular data segment is to represent a single annual layer. However, the determination of such probabilities is challenging in itself, and their calculation will be postponed to chapter 4. They provide the criteria used for determining what should be considered an annual layer, and are of course vital for the methodology. Yet, in the present chapter, only the general framework for annual layer detection will be derived. The aforementioned probabilities, evaluating which segments are the most likely to form an annual layer, are assumed known.

The chapter is very theoretical, and to help the reader a summary of the employed notation can be found in appendix A1.

3.1 An introduction to Bayesian inference

The layer detection algorithm developed is inherently of Bayesian nature. For this reason, an outline of the fundamental concepts in Bayesian inference is given here. These concepts will be applied throughout the rest of the thesis.

In Bayesian statistics, the state of knowledge regarding anything unknown can be described by a probability distribution. Using probabilities as a yardstick of the involved uncertainties, the Bayesian methodology enables statements concerning variables (or unobserved data) to be made when only partial knowledge and uncertain statements are available.

In the following, the quantity to be inferred will be denoted θ, and may comprise a collection of parameters. When wishing to emphasize that we have in mind the entire set of parameters, θ will be referred to as the 'parameter vector', although it does not necessarily fulfill the formal requirements of e.g. linearity for being a physical vector quantity. Likewise, the observed data, y, may sometimes be termed the observation vector. In table 3.1.1 is given an overview of the most common concepts in Bayesian probability theory.

θ	Parameter(s) to be inferred
y	Observed data
$p(\theta)$	Prior probability of θ
$p(\theta\|y)$	Posterior probability of θ (given y)
$L(\theta\|y)$	Likelihood of θ (with fixed y)
$p(u)$	Marginal probability of u
$p(u,v)$	Joint probability of u and v
$p(u\|v)$	Conditional probability of u given v

Table 3.1.1: *The most common probability concepts used in Bayesian inference.*

3.1.1 Bayes' theorem

Using a Bayesian approach, any relevant prior knowledge that we might possess about model parameters is systematically integrated in the analysis by way of prior probabilities, $p(\theta)$. The result of the analysis can be thought of as an update of such prior knowledge based on the observed data, making the methodology ideal for applications in which sequential updates are required. However, as will be discussed shortly, the use of such 'subjective' priors can also be a source of much dispute.

Bayesian statistical conclusions are made in terms of a probability density function $p(\theta|y)$, which is the probability of the parameter θ conditioned on the known data y. This probability distribution is termed the posterior probability density, or just 'the posterior'. Implicitly, the posterior probability density is also conditioned on all further assumptions going into the applied model. The posterior probability distribution can be calculated by use of Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

The term $p(y)$ is the marginal probability distribution of y, i.e. the probability of observing y regardless the value of θ. With the observations fixed, the probability of the involved observations is constant, and $p(y)$ is just a normalization factor:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Thus, the primary task in a Bayesian analysis is to develop a model to compute an estimate of $p(y|\theta)$. Regarding $p(y|\theta)$ as a function of θ (with y fixed), this is termed the

likelihood of θ: $L(\theta|y) = p(y|\theta)$. From the calculated likelihood in combination with our chosen prior, the posterior probabilities can be evaluated.

Bayesian analysis offers a conceptually simple statistical framework with an explicit use of probabilities to quantify the involved uncertainties. The generality of the approach and the ease with which it allows for even very complex models (many parameters, complicated probability distributions etc.) make it applicable to a wide range of statistical problems. In practice, the main limitation of the methodology is often the computational burden associated with calculation of the appropriate likelihood function. However, as a result of the ever-expanding computational power, Bayesian inference is becoming increasingly popular.

3.1.2 Prior probabilities

The prior probability distribution – or simply 'the prior' – reflects the state of knowledge on the parameter values prior to the arrival of any data. The incorporation of prior information is one of the key capabilities of Bayesian analysis. At the same time, however, the subjectivity involved in assessing these priors represents one of the major criticisms of the Bayesian methodology. In many cases, however, a prior can be estimated from other sources, hereby greatly reducing the subjectivity of the choice. If this is not the case, an uninformative prior may be used.

Uninformative and improper priors

An uninformative prior expresses only vague information about a variable. This could e.g. be knowledge regarding its sign or an interval of allowed values, with equal probabilities assigned to each possibility. Given the limited prior knowledge available, such analyses often lead to results very similar to those derived from conventional statistics.

When working with probability densities, however, the assignment of equal probabilities sometimes leads to the use of improper priors: Priors which are not integrable functions, and therefore cannot be probability densities. Such a situation arises e.g. when wishing to assign equal probabilities to all values from $[0, \infty[$. Fortunately, by rewriting Bayes' theorem as:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\theta \; p(y|\theta)p(\theta)d\theta}$$

we see that the prior probabilities $p(\theta)$ need not to be normalized for the posterior to be a sensible probability density, which integrates to one. The prior probabilities just need to be specified in the correct proportion.

Informative priors

If explicit information on the parameter in question is available, an informative prior should be employed. Very often, such knowledge is based on previous experience from similar data, in which case a previous posterior may be used as prior for the current problem. In this way, the prior contains information from previous collected data, and the analysis presents an update of this knowledge based on current data. With an increasing amount of data, the prior will largely become determined by evidence from the data, and not depend on the original choice of prior.

Conjugate priors

Very often, algebraic convenience advocates for the use of conjugate priors [*Raiffa and Schlaifer*, 1961]. A conjugate prior for a given likelihood function will result in a posterior of the same family as the prior distribution. The Gaussian distribution is e.g. self-conjugate: Given a Gaussian likelihood function and choosing a Gaussian prior over its mean, the resulting posterior distribution will also be Gaussian. By using conjugate priors, a closed-form of the resulting posterior can be obtained, hereby avoiding computationally intensive numerical integrations.

Hyper-parameters and hyper-priors

For sequential estimation, in which the resulting posterior probability distribution is to be used as prior for subsequent data analysis, the parameters of the underlying model used for calculating the likelihood function are often not specified directly, but given as probability distributions in terms of their prior probabilities. To avoid confusion, the parameters describing these prior probability distributions are called hyper-parameters.

Conjugate priors are particularly convenient when dealing with sequential estimations, rendering the process of how the likelihood function continuously updates the posterior distribution more straight-forward and intuitive. When using a conjugate prior, the above can simply be described as a change in hyper-parameters due to the information added by the data. The change in hyper-parameters over time can be regarded as the evolution of the system over time.

As an example: A parameter is described by a normal distribution with known variance (σ^2) but unknown mean (μ). The unknown mean of the distribution (μ) may itself be considered a variable, which can be described by a normal distribution with hyper-parameters μ_0 and σ_0, which is a conjugate prior to the normal distribution. As is characteristic for conjugate hyper-parameters, these have a dimensionality one larger than that of the original model parameter. The resulting posterior distribution is again a normal distribution with variance σ^2, and an analytical solution exist for the calculation of hyper-parameters of the posterior based on the values of hyper-parameters of the prior.

A prior probability distribution of a hyper-parameter is called a hyper-prior. In principle, this can be iterated infinitely, allowing for hyper-hyper-parameters (the parameters of a hyper-prior) etc. However, the increased model complexity generally prohibits more than just a few of such iterations.

3.2 Hidden Markov Models

A Hidden Markov Model (HMM) is a stochastic signal model, which can be used for modeling the output of a system displaying Markovian behavior, i.e. a stochastic system which transits between states, and where the next state of the system depends only on the current state. By comparison to observed data, knowledge on the nature of the underlying signal can then be obtained. The concept of Hidden Markov Models was originally introduced in the late 1960s [*L. E. Baum and Petrie*, 1966], and has successfully been applied for pattern recognition in the field of machine speech recognition since the mid-70s

[*Jelinek et al.*, 1975]. A review of the subject and its use for speech recognition is found in *Rabiner* [1989]. Modern general-purpose speech recognition software is almost exclusively based on Hidden Markov Models. However, the rich mathematical structure of Hidden Markov Models can form the theoretical basis for a wide range of signal modeling applications, spanning from magnetic resonance imaging (MRI) brain mapping [*Faisan et al.*, 2005] via electrocardiography (ECG) [*Antti*, 1996; *Thoraval et al.*, 1994] to the analysis of protein structures [*Schmidler et al.*, 2000] as well as financial time series [*Bulla and Bulla*, 2006]. It will here be applied for annual layer recognition in ice core data.

For Hidden Markov Modeling to be applicable, one must be able to unequivocally define a finite number of possible states of the system. The state of the system corresponding to any 'time' t is considered a stochastic variable. The use of t for indexing is owed to Hidden Markov Modeling usually being applied on time series. For the current purpose, t will be used as an index for depth, and hence has nothing to do with the resulting timescale.

The variable describing the state of the system at t will in the following be denoted by S_t. Its outcome is $\ell_j,\ j \in \{1,2,\dots,J\}$, J being the number of possible states of the system. The state of the system is assumed to change stochastically in such a way that the state sequence, $S_{1:T}$, is a Markov chain, i.e. the next state only depends on the current state of the system. In this way, the model only contains limited knowledge on past history.

When a direct outcome of the state sequence can be observed, it is easy to characterize the statistical nature of this signal. However, in many circumstances this is not possible. Instead it may be possible to observe the influence of the state sequence on another stochastic process, the outcome of which is seen as a sequence of observations $\boldsymbol{o}_{1:T}$. This is illustrated in figure 3.2.1. The model hereby has a two-layer structure, with the state sequence providing the unknown 'truth', and each observation by itself only providing incomplete information on the current state. Using a Hidden Markov Model on such data, it may still be possible to infer a statistical estimate of the underlying hidden Markovian state sequence, hence the name of the method.

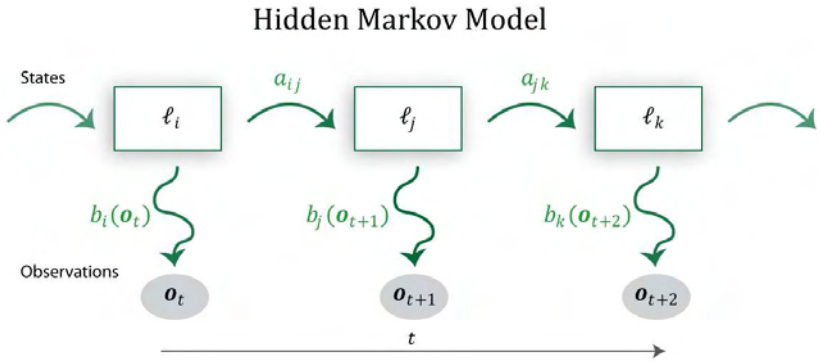


Figure 3.2.1*: The two-layer structure of a Hidden Markov Model (HMM). Two stochastic processes influence the resulting sequence of observations. One is described by the probability of state transitions (a_{ij} etc.), and another is described by the probability of each observation given a specific state ($b_i(\boldsymbol{o}_t)$ etc.). See section 3.3.1 for further explanation of the applied notation.*

When applying a Hidden Markov Model to identify and count annual layers in a paleoclimatic data series, an obvious choice for the states of the system is a labeling with the actual layer number (i.e. year) corresponding to every single data point. We then wish to recover the most likely state sequence giving rise to the observation sequence, which may be any kind of ice core data on a depth scale. This is the desired depth-age relation.

The sequence of encountered years in a data series can be viewed as a Markov chain – a very simple Markov chain, given that the years occur in sequential order without any skipping. However, the actual state sequence, consisting of one annual layer label per data point, is not a Markov chain. For a given depth interval, annual layer thicknesses in ice cores are approximately log-normal distributed [*Andersen et al.*, 2006a]. Hence, from a perspective based only on state sequence probabilities: Assuming the system at a given t to be in state ℓ_i (i.e. data point t is part of layer i), the probability P_{ij} of next being in state ℓ_j (i.e. to be in layer j at data point $t+1$), does not solely depend on the values of i and j (with $P_{ij} = 0$ for $j \neq \{i, i+1\}$). It also depends on the number of data points already encountered in layer i at t, as given by the probability of the resulting annual layer thickness.

Such a state sequence, where the changes in state are endowed with a Markov property, but with holding times of each state distributed according to a specific probability distribution, is called a semi-Markov chain. It can be envisioned as a doubly embedded Markov chain with no self-transitions, which is often a convenient way of representation. In this case, each generalized state $q_j = (\ell_j, d_j)$ includes a duration parameter as well as the actual state label, both of which may depend on the previous generalized state of the system (figure 3.2.2). Allowing the underlying stochastic process to be semi-Markov, this variant of a Hidden Markov Model is in the literature sometimes called a Hidden Semi-Markov Model (HSMM), or – depending on the specific assumptions of the model, its application area and the author – a segment model [*Ostendorf et al.*, 1996; *Yu*, 2010].

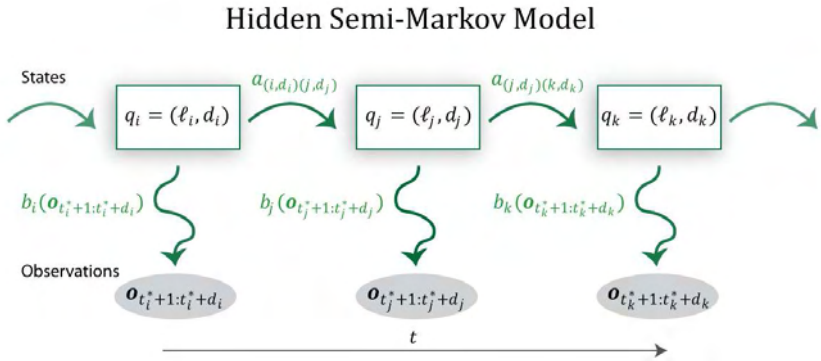


Figure 3.2.2: *A schematic drawing of a Hidden Semi-Markov Model (HSMM). States are now appended with a duration parameter, and observations are collected in segments. State transition probabilities ($a_{(i,d_i)(j,d_j)}$ etc.) as well as observation probabilities ($b_i(\boldsymbol{o}_{t_i^*+1:t_i^*+d_i})$ etc.) may depend both on state labels and state durations. For a further description of the employed notation, see section 3.3.1.*

Within continuous speech recognition, an increasing amount of research in the last decade has gone into extending the applied algorithms to Hidden Semi-Markov Models and segments models [*Ostendorf et al.*, 1996; *Russell and Holmes*, 1997]. Depending on the specific objective of the modeling, the hidden state sequence may in this case either be composed by phones, syllables or entire words, with the size of the vocabulary (i.e. number of possible states) adjusted hereafter. The observed data sequence is composed of an audio recording of a sentence or individual words, which has first passed through some preprocessing. As a speech sound can be characterized by the amount of energy in different frequency bands, the preprocessing is usually done by spectral analysis of the speech signal from which the cepstral components[4] is calculated. Such feature extraction of the signal simplifies the data processing task without discarding too much information from the signal. From the multiple resulting data sequences (each cepstral component of the speech signal is used as a single data series), the most likely combination of words is sought. Via the assumed state transition probabilities of the Hidden Markov process, the recognized utterance may be subjected to syntactic and semantic constraints, hence incorporating 'human' knowledge of sentence constructions. Using the simpler HMMs, no assumption of the duration of a specific speech segment (phone/syllable/word) is made, which implicitly gives rise to a geometric duration distribution. It has been shown that the performance of speech recognition systems improves drastically when instead using HSMMs, which are able to take the proper duration distributions into account [*Gish*, 1993].

The application of Hidden Markov Modeling for annual layer counting in ice core data differs from most other applications of HMMs in the following ways. One is the simplicity of the changes in state, one year simply pursuing the previous. On the other hand, this simplicity is combined with a large variability from one year to the next of how an annual layer is expressed in the data series, which gives rise to a rather challenging pattern recognition problem.

3.3 Overview of layer detection model

An annual layer recognition algorithm based on Hidden Markov Modeling is inherently of Bayesian nature. The resulting annual layer boundaries are given as probability distributions, both in depth and layer number, which are calculated based on a priori knowledge from known state sequence probabilities and updated based on the observed data, hereby forming the posterior probabilities. It is the repeated application of Bayes' theorem that leads to the final layer detection algorithm.

Consider the option of envisaging all possible segmentations of an observation sequence, one at a time, and calculating the respective probabilities. In this way, the most likely segmentation can be found, and the result will be based on the entire observation sequence in consideration. In reality, however, such an approach is not feasible. Fortunately, the

[4] The power cepstrum is the inverse Fourier transform of the logarithm of the power spectrum of a signal [Norton, M. P., and D. G. Karczub (2003), *Fundamentals of noise and vibration analysis for engineers*, 2nd ed., 631 pp., Cambridge University Press, Cambridge, New Yorkibid.].

same probabilities can be efficiently calculated by recursion by use of one of two algorithms: The Forward-Backward algorithm (section 3.4) or the Viterbi algorithm (section 3.6). The difference between the two lies in the specific definition of a most likely layering. The approach here is primarily focused on probabilities calculated using the Forward-Backward algorithm, but results from the Viterbi algorithm will be considered as well.

When calculating these probabilities, an entire observation sequence is taken into account at once. The likelihood of a given observation segment representing an annual layer is therefore judged not only by its own resemblance to an annual layer, but is seen in conjunction with the likelihood of the proposed annual layers and annual layer thicknesses on either side. In this way, the algorithm works very similar to what is implicitly done by eye when counting layers manually.

The name "Forward-Backward algorithm" is derived from the way the algorithm makes such judgment in a rigorous yet efficient manner by executing respectively a forward pass of the data series, which contains the information included in all previous data, and a backward pass, containing information from all subsequent data. In this way, the entire data sequence is used for inferring the most likely layering and the involved uncertainties. The derived uncertainties provide an estimate of the counting error in much the same way as is done manually.

The algorithm is able to use the data itself to improve on estimates of the parameter values describing e.g. the annual layer template (unsupervised learning). In other words, the algorithm is able to use the information from distinct layers in the entire sequence to deal with sections herein of less obvious layering. This gives rise to high performance of the algorithm, even when only imperfect knowledge is available on employed parameter values. The algorithm is therefore robust against gradual changes with depth in annual layer appearance, changes in layer thickness distributions etc. Such robustness is necessary as the layer thicknesses and appearance highly depends on the climate regime during deposition as well as ice-flow induced thinning of layers with depth.

Also many other features of this annual layer counting algorithm based on Hidden Markov Modeling bear similarities to a manual approach. Sections containing missing data are e.g. treated much the same way: The most likely number of annual layers contained within such a section is estimated based on the assumed annual layer thickness distribution along with any possible layer fragments on either side.

Finally, the method can relatively easily be extended to allow the incorporation of multiple data series containing an annual layer signal, and infer the most likely annual layering based on all of these data series at once. It therefore provides the necessary statistical framework to allow for an automated multiple-parameter counting algorithm. For the current purpose, this property of the algorithm has been used to incorporate information from derivatives of the observation sequence as well as the observation sequence itself.

Features, such as multi-parameter counting and the treatment of an entire observation sequence at once, have otherwise proved hard to incorporate in automatic methods. Consequently, manual annual layer counting in ice core data have hitherto generally provided the best results.

3.3.1 Notation

In the following, the potential annual layers encountered during the observation sequence will be denoted by $\ell_j \in \mathcal{L} = \{\ell_1, \ell_2, \dots, \ell_J\}$. The total observation sequence of length T will be represented as $\boldsymbol{o}_{1:T}$. Equivalently, a section of the observation sequence covering data between t_1 and t_2 (both included) will be represented as $\boldsymbol{o}_{t_1:t_2}$. Layer durations, d, are given in terms of number of observations covered by each layer. These are allowed to assume integer values in the finite set $\mathcal{D} = \{1,2, \dots, D\}$ with a prescribed probability distribution $p(d)$.

The hidden state of the system corresponding to observation t will be denoted by S_t. The sequence $S_{1:T}$ is thus a sequence of stochastic variables, whose values are confined to the set $\mathcal{L}$. A realization of such a sequence, corresponding to observations $\boldsymbol{o}_{t_1:t_2}$, will be written $s_{t_1:t_2}$. Following *Yu* [2010], the short hand notation $S_{[t_1:t_2]} = \ell_j$ will in the following be used to signify that layer j starts exactly at t_1 and ends exactly at t_2 (both data points included). The notation $S_{[t} = \ell_j$ indicates that layer j starts at t, while containing no information on where the layer ends. Similarly, $S_{t]} = \ell_j$ signifies that layer j ends at t, but says nothing about where the layer started. $S_t = \ell_j$ only implies that observation t is a part of layer j, and bears no information on the state of the surrounding observations.

The probability of a state transition from the generalized state (ℓ_i, d') to state (ℓ_j, d) is assumed stationary in time. In a formal sense, and using the above definitions, this transition probability is defined by:

$$a_{(i,d')(j,d)} \equiv P\left(S_{[t+1:t+d]} = \ell_j | S_{[t-d'+1:t]} = \ell_i\right)$$

More specifically, it is the probability of entering layer j having duration d, provided that the previous layer i of duration d' has just ended. However, the annual layers follow each other in a sequential manner, and are assumed to share the same layer thickness probability density function. If furthermore the duration of the new layer ℓ_j is assumed independent on both numbering and duration of the previous layer, this transition probability simplifies to:

$$a_{(i,d')(j,d)} = a_{ij} p(d),$$

Previous studies point to the existence of a slightly negative correlation between successive layer thicknesses [*Fisher et al.*, 1985]. Such correlation has not been taken into account when using the above formulation.

With $p(d)$ being the layer thicknesses probability distribution, and a_{ij} being the probability of a transition from layer ℓ_i to layer ℓ_j (regardless of layer thicknesses), we have:

$$a_{ij} \equiv P\left(S_{[t+1} = \ell_j | S_{t]} = \ell_i\right) = \delta_{i,j-1} = \begin{cases} 1, & j = i+1 \\ 0, & otherwise \end{cases}$$

In the last part, the Kronecker delta notation $\delta_{i,j}$ has been employed. Note, that the state transition probabilities are only concerned with the probability of the resulting state sequence, and independent of the actual observations.

The dependency on the observations is contained in the emission probabilities. The emission probability is defined as the conditional probability of observing a given sequence of observations, when assuming these to form an annual layer ℓ_j:

$$b_j(\boldsymbol{o}_{t+1:t+d}) \equiv P\left(\boldsymbol{o}_{t+1:t+d} \mid S_{[t+1:t+d]} = \ell_j\right) \tag{3.3.1}$$

The observation sequence $\boldsymbol{o}_{1:T}$ is allowed to be a sequence of vector observations, meaning that several observations may be connected to each index t. In case of the ice core data, this implies that the annual layer counting method can be extended to a full multi-parameter counting approach, with $\boldsymbol{o}_t$ being a vector containing the entire collection of chemistry data measured at depth t.

All annual layers are to be described as a product of the same seasonal deposition process. For this reason, the probabilities $b_j(\boldsymbol{o}_{t+1:t+d})$ are independent of the specific layer under consideration, and when considering the present task of annual layer detection, the dependence of the emission probabilities on j can be left out.

The calculation of these emission probabilities is the very heart of the layer detection algorithm. The emission probabilities evaluate which observation segments resemble an annual layer and which do not. The performance of the layer detection algorithm crucially depends on a proper description of how an annual layer is being represented by the observations. All model parameters used for calculating the emission probabilities, along with those used in the parameterization of $p(d)$, are collected in the variable θ, which will be used as short-hand notation in the following. Thus, θ contains all free parameters of the Hidden Markov Model. First, however, consider the probabilities $b(\boldsymbol{o}_{t+1:t+d})$ known from data. The calculation of these will be dealt with in chapter 4.

3.4 The Forward-Backward algorithm

In the terminology of a HMM, determining the best annual layering in a section of observed ice core data corresponds to inferring the most likely hidden state sequence giving rise to the observed data. For the Forward-Backward algorithm, this is to be understood as the state sequence in which each state individually has maximum posterior probability, when conditioned on the entire observation sequence. Using this definition, the most likely state at t is the state $\ell_j \in \mathcal{L}$ satisfying:

$$\ell_{\mathrm{MAP}}(t) = \underset{\ell_j}{\operatorname{argmax}} \left\{P\left(S_t = \ell_j \mid \boldsymbol{o}_{1:T}, \theta\right)\right\} \tag{3.4.1}$$

The dependence on the applied model parameters (θ) is included to clarify that the result depends on the chosen model and model parameters used for describing annual layers and their thickness probability density function. The expression ‘argmax’ stands for the argument that leads to the maximum, and MAP is short for Maximum a Posteriori, i.e. the maximum of the posterior probability function. The Forward-Backward algorithm presents a way in which such posterior probabilities can be evaluated in a rigorous, yet efficient way.

Other definitions of a "most likely state sequence" are possible as well. The one most commonly used is that of the Viterbi path [*Viterbi*, 1967], discussed in section 3.6. Indeed, the definition in (3.4.1) may lead to state sequences that are not allowed. This could e.g. be a layer sequence in conflict with the assumption of layers being laid down in successive order. However, the main objective of the present analysis is not an ideal segmentation of the observation sequence into individual annual layers. Rather, it is a best estimate of the total number of annual layers within any given depth interval. Using the above definition, it is also possible to calculate the entire probability distribution for each state variable $S_t,\ t \in \{1, \dots, T\}$, when taken into account all of the observations $\boldsymbol{o}_{1:T}$. Hereby not only the most likely annual layering is inferred, concurrently also the uncertainty of the resulting layering is estimated.

In principle, such probabilities (3.4.1) can be calculated by brute force by considering all possible state sequences one at a time, calculating their respective probabilities, and adding up those which give rise to state ℓ_j at t. However, even if constricting ourselves to applications of short observations sequences with a small number of possible states and durations, such a calculation is an overwhelming undertaking, which quickly creates a heavy computational burden.

A much more efficient way to calculate these probabilities is by means of first calculating the joint probability of ending layer ℓ_j of duration d at t and observing the observation sequence $\boldsymbol{o}_{1:T}$:

$$\eta_t(j,d) \equiv P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:T} | \theta\right) \tag{3.4.2}$$

Although by first sight not appearing any simpler than the previous equation (3.4.1), these probabilities can be calculated recursively using a generalized version of the Forward-Backward algorithm commonly used in HMMs, and extended to the appliance for HSMMs.

The joint probability of a collection of events can be calculated by multiplying the probabilities of each event conditioned on all other events: $P(A,B,C) = P(A) \cdot P(B|A) \cdot P(C|A,B)$. Simplifying notation by skipping the dependence on the model parameters θ throughout the subsequent derivations, equation (3.4.2) can therefore be rewritten as:

$$\begin{aligned} P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:T}\right) &= P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}, \boldsymbol{o}_{t+1:T}\right) \\ &= P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right) \cdot P\left(\boldsymbol{o}_{t+1:T} \middle| S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right) \\ &= P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right) \cdot P\left(\boldsymbol{o}_{t+1:T} \middle| S_{[t-d+1:t]} = \ell_j\right) \end{aligned}$$

The last equality rests upon the assumed independence of the observations $\boldsymbol{o}_{t+1:T}$ on $\boldsymbol{o}_{1:t}$, which is an assumption inherent to the HMM approach.

By doing so, the initial problem has been substituted by that of calculating the forward and backward variables $\alpha_t(j,d)$ and $\beta_t(j,d)$ defined as:

$$\begin{aligned} \alpha_t(j,d) &\equiv P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right) \\ \beta_t(j,d) &\equiv P\left(\boldsymbol{o}_{t+1:T} | S_{[t-d+1:t]} = \ell_j\right) \end{aligned} \tag{3.4.3}$$

In terms of these two variables, we have the following identity:

$$\eta_t(j,d) = \alpha_t(j,d) \cdot \beta_t(j,d)$$

As it will be shown, both $\alpha_t(j,d)$ and $\beta_t(j,d)$ can be calculated in an efficient manner by recursion. The number of calculations required for a brute force approach is exponentially increasing with T, J and D. By way of recursion, the order of computational complexity is reduced to being linear in these variables.

The equation above can be interpreted as follows: In order to find the probability of a specific layer (j,d) ending at t, both observations before and after t must be taken into account. In the Forward-Backward algorithm, this is done by a double pass of the observation sequence. The forward pass takes account of the entire observation sequence up to t, while the backward pass takes account of information based on later observations. The best estimate of the hidden state sequence is then found by combining the two.

3.4.1 Forward message pass

The forward variable $\alpha_t(j,d)$ gives the joint probability of ending state ℓ_j with duration d at t, and of observing the partial observation sequence $\boldsymbol{o}_{1:t}$ (3.4.3). In the general case, $\alpha_t(j,d)$ can be calculated recursively by [*Yu*, 2010]:

$$\begin{aligned}\alpha_t(j,d) &= P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right) \\ &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in \mathcal{D}} \alpha_{t-d}(i,d') a_{(i,d')(j,d)} b_j(\boldsymbol{o}_{t-d+1:t})\end{aligned} \tag{3.4.4}$$

A derivation of this equality is found in box 1. (This is a general equation, but I have not been able to find its derivation anywhere.)

With the simplifying assumptions pertinent to the present case (sequential states, duration of the next layer independent on duration and number of the previous layer, layer signal independent on layer number), it can be reduced to:

$$\begin{aligned}\alpha_t(j,d) &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} \alpha_{t-d}(i,d')\, \delta_{i,j-1}\, p(d)\, b(\boldsymbol{o}_{t-d+1:t}) \\ &= \sum_{d' \in D} \alpha_{t-d}(j-1,d')\, p(d)\, b(\boldsymbol{o}_{t-d+1:t}) \\ &= p(d) b(\boldsymbol{o}_{t-d+1:t}) \sum_{d' \in D} \alpha_{t-d}(j-1,d') \\ &= p(d)\, b(\boldsymbol{o}_{t-d+1:t})\, \tilde{\alpha}_{t-d}(j-1)\end{aligned} \tag{3.4.5}$$

A new variable $\tilde{\alpha}_t(j)$ has here been introduced to simplify notation. It is the total probability of ending layer j at t, while observing the partial observation sequence $\boldsymbol{o}_{1:t}$:

$$\tilde{\alpha}_t(j) \equiv P\left(S_{t]} = \ell_j, \boldsymbol{o}_{1:t}\right) = \sum_{d \in \mathcal{D}} \alpha_t(j,d)$$

Before evaluating the recursion, the initialization conditions must be considered. The most common initialization assumption is that the first state begins at $t = 1$ (leading to $\tilde{\alpha}_{t\leq 0}(j) = 0$ for all j), although sometimes a more general assumption of unknown starting position of the first state somewhere before beginning of the observation sequence is used. In this case [Yu, 2010]:

Box 1: Recursive formula for the forward variable

We here derive the recursive equation for $\alpha_t(j,d)$ for the general case where the probability of successive states depends both on the individual states and their duration. To simplify the notation, the conditioning on the model parameters (θ) will not be explicitly annotated. We have:

$$\begin{aligned}\alpha_t(j,d) &\equiv P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right)\\ &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} P\left(S_{[t-d-d'+1:t-d]} = \ell_i, S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right)\\ &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} P\left(S_{[t-d-d'+1:t-d]} = \ell_i, S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t-d}, \boldsymbol{o}_{t-d+1:t}\right)\\ &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{t-d+1:t} | S_{[t-d-d'+1:t-d]} = \ell_i, \boldsymbol{o}_{1:t-d}\right) \cdot\\ &\qquad P\left(S_{[t-d-d'+1:t-d]} = \ell_i, \boldsymbol{o}_{1:t-d}\right)\end{aligned}$$

Utilizing the two-level structure of the Hidden Markov Model, it can be realized that knowledge of the underlying state sequence provides all necessary information to evaluate the probabilities of subsequent states as well as observations. Hence, given that state ℓ_i is assumed to end at time $t-d$, the conditioning on $\boldsymbol{o}_{1:t-d}$ in the first term above can be dropped. The last term can be recognized as $\alpha_{t-d}(i,d')$:

$$\begin{aligned}\alpha_t(j,d) &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{t-d+1:t} | S_{[t-d-d'+1:t-d]} = \ell_i\right) \cdot \alpha_{t-d}(i,d')\\ &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} P\left(\boldsymbol{o}_{t-d+1:t} | S_{[t-d+1:t]} = \ell_j, S_{[t-d-d'+1:t-d]} = \ell_i\right) \cdot P(S_{[t-d+1:t]}\\ &\qquad = \ell_j | S_{[t-d-d'+1:t-d]} = \ell_i) \cdot \alpha_{t-d}(i,d')\\ &= \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} P\left(\boldsymbol{o}_{t-d+1:t} | S_{[t-d+1:t]} = \ell_j\right) \cdot P(S_{[t-d+1:t]} = \ell_j | S_{[t-d-d'+1:t-d]} = \ell_i) \cdot \alpha_{t-d}(i,d')\end{aligned}$$

In the last equality we have again made use of the assumption that the probability of a given observation sequence $\boldsymbol{o}_{t_1:t_2}$is fully described by the state of the system between t_1 and t_2. Recognizing the variables $b_j(\boldsymbol{o}_{t-d+1:t})$ and $a_{(i,d')(j,d)}$ in the above expression, we finally arrive at the following recursive formula for $\alpha_t(j,d)$:

$$\alpha_t(j,d) = \sum_{\ell_i \epsilon \mathcal{L}\setminus\{\ell_j\}} \sum_{d' \in D} b_j(\boldsymbol{o}_{t-d+1:t})\, a_{(i,d')(j,d)} \alpha_{t-d}(i,d')$$

(3.4.6) $$\tilde{\alpha}_{t\leq 0}(j) = 1, \quad j \in \{0, \dots J\}$$

For the layer detection algorithm developed here, these initialization conditions can be improved. The simple structure of this particular Hidden Markov Model allows us to utilize any available information about the ending position of the layer prior to start of the observation sequence. Without any loss of generality, this layer will be taken as 'layer 0' (ℓ_0), and all subsequent layers are counted upwards from this. Thus the first observation is always a part of ℓ_1. The initial distribution of $\tilde{\alpha}_t(j)$ for $t \leq 0$ is given by:

(3.4.7) $$\pi_t(j) \equiv \tilde{\alpha}_{t\leq 0}(j) = \begin{cases} P\left(S_{t]} = \ell_0\right), & j = 0 \\ 0, & j \neq 0 \end{cases}$$

The notation $\pi_t(j)$ has here been used to indicate the initial conditions of the Forward-Backward algorithm. The above initialization condition can be obtained by allowing the

observations outside the sampling period to take on any possible value, such that the probability of encountering any one of these is equal to 1.

However, it turned out that due to the flexibility of the Forward-Backward algorithm to account for the variation in thickness of individual layers, the algorithm is not very strongly dependent on these.

Provided that $b(\boldsymbol{o}_{t-d+1:t})$ and $p(d)$ are known, $\alpha_t(j,d)$ can now be calculated for all t, j and d.

3.4.2 Backward message pass

The backward variable $\beta_t(j,d)$ is defined as:

$$\beta_t(j,d) = P\left(\boldsymbol{o}_{t+1:T} | S_{[t-d+1:t]} = \ell_j\right)$$

Whereas the forward variable takes care of information from the first part of the observation sequence, the backward variable includes information contained in the second part of the observation sequence. It can also be calculated recursively [*Yu*, 2010] (see box 2 for a derivation):

$$\beta_t(j,d) = P\left(\boldsymbol{o}_{t+1:T} | S_{[t-d+1:t]} = \ell_j\right) = \sum_{\ell_i \in \mathcal{L} \setminus \{\ell_j\}} \sum_{d' \in D} a_{(j,d)(i,d')} b_i(\boldsymbol{o}_{t+1:t+d'}) \beta_{t+d'}(i,d')$$

The same simplifications as those used for the forward pass lead to the reduced equation:

$$\beta_t(j,d) = \sum_{d' \in D} p(d') b(\boldsymbol{o}_{t+1:t+d'}) \beta_{t+d'}(j+1,d') \tag{3.4.8}$$

No assumptions are made regarding the end of the last layer. Hence, the backward pass is initialized using the general assumption that the last layer ends somewhere after the last observation in the observation sequence. Given the definition of $\beta_t(j,d)$, and assuming the observations after T to take on any possible value, such initialization condition for $\beta_t(j,d)$ implies that:

$$\beta_{t \geq T}(j,d) = 1 \tag{3.4.9}$$

From the independence of this initial condition both on j and d, and the non-existence of these two parameters in the recursion formula for $\beta_t(j,d)$ (3.4.8), it is seen that the backward variable only depends on t:

$$\beta_t = \sum_{\ell_j \in \mathcal{L}} P\left(\boldsymbol{o}_{t+1:T} | S_{t]} = \ell_j\right) = \sum_{d' \in D} p(d') b(\boldsymbol{o}_{t+1:t+d'}) \, \beta_{t+d'}$$

Given the definition of $\beta_t(j,d)$, and the general structure of the assumed model, the independence on j and d should not be a major surprise: As each layer is assumed independent on the previous layer, the knowledge that a layer just ended is the only information needed for calculating the probability of the remaining observation sequence. All other properties of this layer are irrelevant.

Box 2: Recursive formula for the backward variable

Here, the general recursive equation for efficient calculation of the backward variable $\beta_t(j,d)$ will be derived. As before, the conditioning on the model and corresponding parameters (θ) will not be explicitly annotated. Manipulating the definition for the backward variable, we find that:

$$\begin{aligned}\beta_t(j,d) &\equiv P(\boldsymbol{o}_{t+1:T}|S_{[t-d+1:t]}=\ell_j)\\ &= \sum_{\ell_i\epsilon\mathcal{L}\setminus\{\ell_j\}}\sum_{d'\in D} P(\boldsymbol{o}_{t+1:t+d'},\boldsymbol{o}_{t+d'+1:T},S_{[t+1:t+d']}=\ell_i|S_{[t-d+1:t]}=\ell_j)\\ &= \sum_{\ell_i\epsilon\mathcal{L}\setminus\{\ell_j\}}\sum_{d'\in D} P(\boldsymbol{o}_{t+d'+1:T}\,|\boldsymbol{o}_{t+1:t+d'},S_{[t+1:t+d']}=\ell_i,S_{[t-d+1:t]}=\ell_j)\\ &\qquad\cdot P\big(\boldsymbol{o}_{t+1:t+d'},S_{[t+1:t+d']}=\ell_i|\,S_{[t-d+1:t]}=\ell_j\big)\end{aligned}$$

Making use of the Markovian property of the state sequence, as well as observations only being dependent on the state sequence, the conditioning in the first term can be reduced to that of $S_{[t+1:t+d']}=\ell_i$. Hence, the equation can be rewritten as:

$$\beta_t(j,d)=\sum_{\ell_i\epsilon\mathcal{L}\setminus\{\ell_j\}}\sum_{d'\in D} P(\boldsymbol{o}_{t+d'+1:T}\,|S_{[t+1:t+d']}=\ell_i)\cdot$$
$$P(\boldsymbol{o}_{t+1:t+d'}\,|\,S_{[t+1:t+d']}=\ell_i,S_{[t-d+1:t]}=\ell_j)\cdot P\big(S_{[t+1:t+d']}=\ell_i|\,S_{[t-d+1:t]}=\ell_j\big)$$

Also here, the conditioning on $S_{[t+1:t+d']}=\ell_i$ can be dropped from the middle component of the above equation. Recognizing then the expressions for $\beta_{t+d'}(i,d')$, $b_i(\boldsymbol{o}_{t+1:t+d'})$ and $a_{(j,d)(i,d')}$, we arrive at:

$$\begin{aligned}\beta_t(j,d) &= \sum_{\ell_i\epsilon\mathcal{L}\setminus\{\ell_j\}}\sum_{d'\in D} P(\boldsymbol{o}_{t+d'+1:T}\,|S_{[t+1:t+d']}=\ell_i)\;\cdot\\ &\qquad P(\boldsymbol{o}_{t+1:t+d'}\,|\,S_{[t+1:t+d']}=\ell_i)\,P\big(S_{[t+1:t+d']}=\ell_i|\,S_{[t-d+1:t]}=\ell_j\big)\\ &= \sum_{\ell_i\epsilon\mathcal{L}\setminus\{\ell_j\}}\sum_{d'\in D}\beta_{t+d'}(i,d')\,b_i(\boldsymbol{o}_{t+1:t+d'})\,a_{(j,d)(i,d')}\end{aligned}$$

3.4.3 Posterior probabilities of layer positions

Let's now return to the original question: We have an observation sequence $\boldsymbol{o}_{1:T}$. What is the posterior probability of being in layer j at any given index t, i.e. $P(S_t=\ell_j|\boldsymbol{o}_{1:T})$? Consider the joint probability $\gamma_t(j)$ defined by:

$$\gamma_t(j)=P(S_t=\ell_j,\boldsymbol{o}_{1:T})$$

Observe that this joint probability measure only quantifies the probability of being in layer j at t. It does not specify where the layer starts or ends, nor its duration.

The posterior probabilities $\gamma_t(j)$ can be calculated based on the forward and backward variables. The results give the entire probability distributions corresponding to the layer at a given index t. These probability distributions can be used not only for determining the most likely layer corresponding to the observation at t, but also to infer the uncertainties associated with the determination of this most likely layer.

Recall that the probability of ending layer j with duration d at index t, when given the entire observation sequence $\boldsymbol{o}_{1:T}$, is:

$$\eta_t(j,d)\equiv P\big(S_{[t-d+1:t]}=\ell_j,\boldsymbol{o}_{1:T}\big)=\alpha_t(j,d)\beta_t$$

Its marginalized probability distribution is the probability of ending a given layer regardless of its duration:

$$\tilde{\eta}_t(j) \equiv P\big(S_{t]} = \ell_j, \boldsymbol{o}_{1:T}\big) = \sum_{d \in \mathcal{D}} \eta_t(j,d)$$

By definition, the first observation in the observation sequence always belongs to layer 1. Thus, the probability of still being in layer 1 at index t is just the probability of not yet having ended the layer, i.e. the probability of ending layer 1 at or after t. That is:

$$\gamma_t(1) = P(S_t = \ell_1, \boldsymbol{o}_{1:T}) = \sum_{\tau \geq t} P(S_{\tau]} = \ell_1, \boldsymbol{o}_{1:T}) = \sum_{\tau \geq t} \tilde{\eta}_\tau(1)$$

Using this as the initialization condition, the probabilities of the remaining layers can now be calculated recursively. The probability of being in layer j at t equals the probability of being in layer j at $t-1$, minus the probability of having ended layer j at $t-1$, plus the probability of beginning layer j at t (i.e. ending layer $j-1$ at time $t-1$):

$$\gamma_t(j) = \gamma_{t-1}(j) - \tilde{\eta}_{t-1}(j) + \tilde{\eta}_{t-1}(j-1)$$

This recursive formulation gives the probability of being in any layer j at any point in the observation sequence. The maximum a posteriori (MAP) estimate of the annual layer corresponding to the observation at a given t can be found as:

$$\ell_{MAP}(t) \equiv \underset{\ell_j}{\operatorname{argmax}} \left\{P(S_t = \ell_j, \boldsymbol{o}_{1:T})\right\} = \underset{j}{\operatorname{argmax}} \{\gamma_t(j)\}$$

This is the most likely layer at observation t. By considering the width of the derived probability distributions, the uncertainties involved in the inference of the most likely layer can be estimated.

The single most likely layer sequence can be found as the sequence of maximum a posteriori layers for each t. However, this MAP state sequence is not necessarily regular. The maximum a posteriori criterion provides the most likely layer at each t, but this is being determined separately for each t. This means that in sections where the layering is vague, some layer boundaries may correspond to a shift of two layers, or a previous layer may reappear. The MAP layer boundaries can be found as those locations where a change in ℓ_{MAP} occurs. But the most likely total number of layers in a section, as determined by $\ell_{MAP}(T)$, may in principle differ from the number of MAP layer boundaries found. This does not imply that there is anything wrong with the algorithm. It is just a consequence of the employed definition of a best state sequence as the one which individually maximizes the posterior probability of each state. As it will be discussed further in section 3.6, this is a strength rather than a weakness of the layer counting algorithm, as the goal is to produce the best possible chronology down the ice core.

This is how the Forward-Backward algorithm is able to compute the most likely state sequence based on the entire observed data series: The information obtained from respectively a forward and a backward pass of the observation sequence is combined. We will now consider the output of the algorithm when applying it to a small section of visual stratigraphy data from NGRIP.

3.4.4 Output from the Forward-Backward algorithm

In figure 3.4.1A&B, the log-probabilities resulting from respectively a forward and a backward message pass of the Forward-Backward algorithm are shown. The employed data series (figure 3.4.1E) is the visual stratigraphy data from the NGRIP ice core from a depth of 2233 m (age: 47 ka BP).

For the forward variable $\tilde{\alpha}_t(j)$, the probability distribution corresponding to each layer is drawn in different colors. With increasing distance from the beginning of the data series, the $\tilde{\alpha}_t(j)$-values generally decrease, whereas β_t-values generally increase. This decrease in probability with distance from the initiation point of each pass is due to the definition of $\tilde{\alpha}_t(j)$ and β_t as the joint probability of ending a layer at a given position, and observing an ever-increasing number of observations. Due to the log-scale, even small bumps in the log-probabilities correspond to large differences in probability.

By combining the probabilities resulting from the forward and the backward pass of the data sequence, the probabilities $\eta_t(j)$ can be computed (figure 3.4.1C). The probabilities $\eta_t(j)$ estimate the probabilities of ending a given layer j at t. The protruding peaks are therefore locations, as inferred by the algorithm, with high probability of being layer boundaries. The total probability of ending any layer at a given location t can be calculated as the sum of all contributions from the individual layers. A slight decrease in peak height with distance is therefore due to an increase in uncertainty as to which layer the boundary belongs.

The probabilities $\eta_t(j)$ can subsequently be converted to $\gamma_t(j)$, expressing the probability of being in a given layer j at t (figure 3.4.1D). From these values, the most likely layer at a given position can be found, as well as the probability of other layers at that location. Assuming the observation sequence to always start in layer 1, the probability of being in this layer is equal to one at the start of the observation sequence. The further away from the start, the less certainly can the total number of layers be determined. As a result, the maximum probabilities slowly decrease.

From the data employed here, it is seen that the algorithm struggles with the layer around 2233.1 m, where the peak values of $\gamma_t(j)$ suddenly decrease. Indeed, the corresponding peaks in the visual stratigraphy data at this location do look a bit strange, hence explaining the behavior of the algorithm. Nevertheless, the layer detection algorithm manages to recover much the same positions for the annual layer boundaries as those in the GICC05 chronology. The resulting layer boundaries based on the MAP criteria are shown in figure 3.4.1E.

3.4.5 Likelihood of applied model parameters

In addition to inferring the most likely layer boundaries, the Forward-Backward algorithm is able to evaluate the likelihood of the applied model parameters based on the observation sequence. This feature of the algorithm gives the opportunity to improve the model parameter estimates used as input to the algorithm. The calculation of the likelihood of the applied model parameters is shown here, and in chapter 4, a procedure for improving the model parameter estimates will be developed.

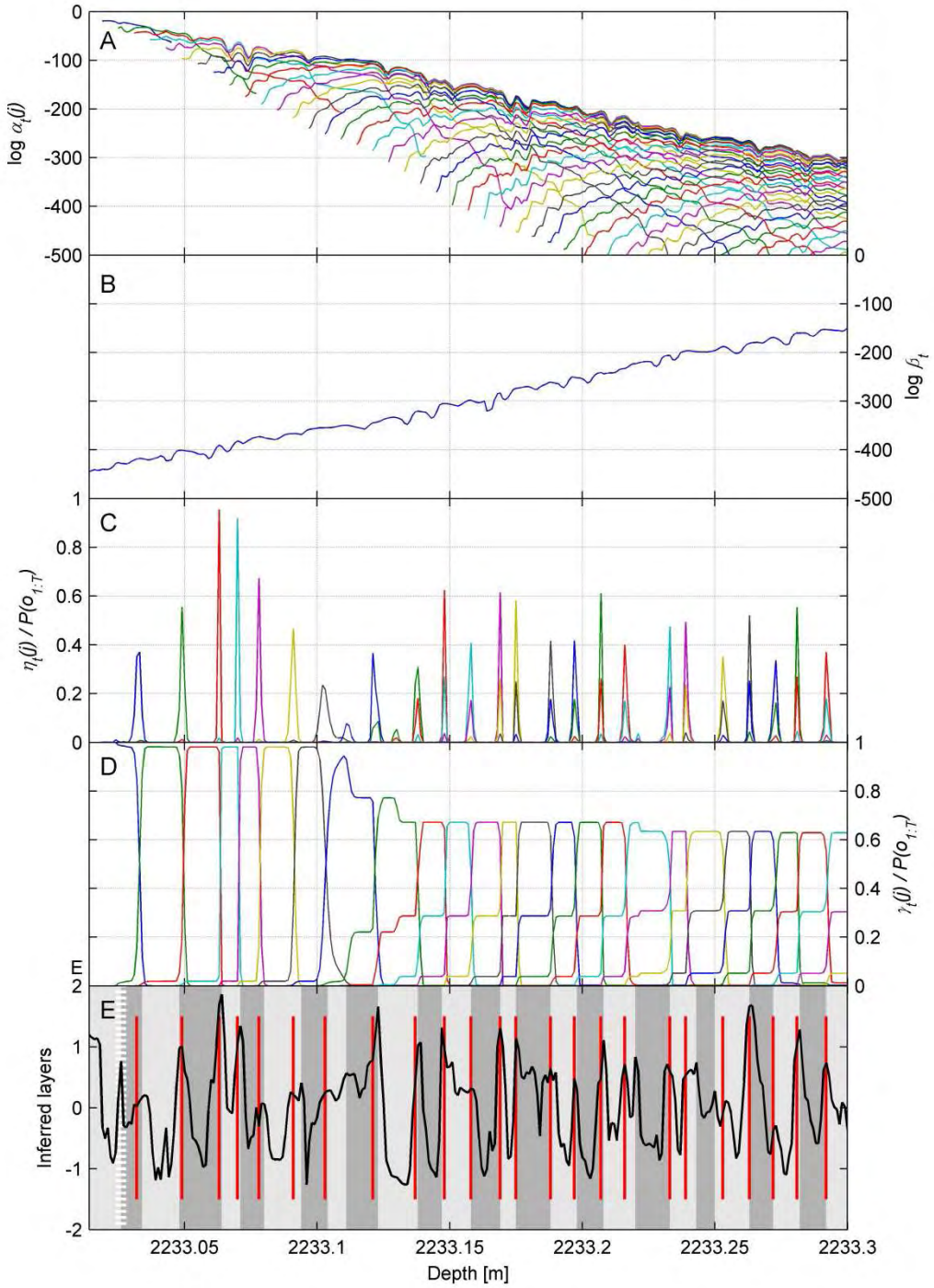


***Figure 3.4.1**: The output from a run of the Forward-Backward algorithm. The data employed (E) are visual stratigraphy data from NGRIP. The layer definition employed is described in chapter 4. The result from the forward pass ($\tilde{\alpha}_t(j)$) and backward pass (β_t) (A,B) combines to form $\eta_t(j)$ and $\gamma_t(j)$ (C,D), which tells about the probability of respectively ending and being in a given layer j at a given position t. Note, that $\tilde{\alpha}_t(j)$ and β_t are shown on log-scale, whereas $\eta_t(j)$ and $\gamma_t(j)$ are not. The resulting MAP layer positions are shown in E. The bright and dark banding shows the annual layers in the GICC05 chronology, uncertain layer boundaries are marked with small horizontal white stripes.*

In a formal sense, the computed posterior probabilities $\eta_t(j)$ and $\gamma_t(j)$ are conditioned on the chosen model and model parameters used for specifying the annual layer characteristics. This set of model parameters is denoted θ.

The probability of observing exactly the current observation sequence is just a number, the value of which can e.g. be calculated by:

$$P(\boldsymbol{o}_{1:T}) = \sum_{\ell_j \in \mathcal{L}} \gamma_t(j) \tag{3.4.10}$$

That this equation holds true for every value of t can be realized from the fact that the total probability of being in any possible layer j at t must equal 1:

$$\sum_{\ell_j \in \mathcal{L}} \gamma_t(j) = \sum_{\ell_j \in \mathcal{L}} P\big(S_t = \ell_j, \boldsymbol{o}_{1:T}\big) = P(\boldsymbol{o}_{1:T}) \sum_{\ell_j \in \mathcal{L}} P\big(S_t = \ell_j | \, \boldsymbol{o}_{1:T}\big) = P(\boldsymbol{o}_{1:T})$$

As a result, the probability of being in layer j at t can be evaluated directly without including the observation sequence probability. A normalized version of $\gamma_t(j)$ can therefore be defined as:

$$\bar{\gamma}_t(j) \equiv P\big(S_t = \ell_j \,\big|\boldsymbol{o}_{1:T}\big) = \frac{P\big(S_t = \ell_j, \boldsymbol{o}_{1:T}\big)}{P(\boldsymbol{o}_{1:T})} = \frac{\gamma_t(j)}{P(\boldsymbol{o}_{1:T})}$$

And likewise for $\eta_t(t)$:

$$\bar{\eta}_t(j,d) \equiv P\big(S_{[t-d+1:t]} = \ell_j | \boldsymbol{o}_{1:T}\big) = \frac{\eta_t(j,d)}{P(\boldsymbol{o}_{1:T})}$$

These expressions will e.g. be used in chapter 5.

For a fixed observation sequence, $P(\boldsymbol{o}_{1:T})$ is constant. To determine e.g. the maximum a posteriori estimate of the layer at t, it has no impact whether or not division with this number has taken place. However, the value of $P(\boldsymbol{o}_{1:T})$ is minuscule, as any specific observation sequence is indeed very unlikely to occur. By having eliminated the extremely small probabilities associated with $P(\boldsymbol{o}_{1:T})$, the normalized probabilities $\bar{\gamma}_t(j)$ and $\bar{\eta}_t(j,d)$ are much easier to interpret.

However, the true power of (3.4.10) is more profound: Throughout the derivations of the posterior probabilities, and therefore also in (3.4.10), we left out the dependence on the model parameters θ describing how an annual layer is expected to appear in the observations. By leaving it in, it can be seen that by calculating the probability of the observation sequence, the likelihood of the chosen model parameters has been inferred:

$$L(\theta|\boldsymbol{o}_{1:T}) = P(\boldsymbol{o}_{1:T}|\theta)$$

Consequently, the Forward-Backward algorithm gives an opportunity to evaluate the likelihood of the employed set of model parameters, hereby presenting us with a method by which the model parameters best suited for modeling the observations can be selected. In other words, a learning process can be implemented. The opportunity of such training of the annual layer detection algorithm is a major advantage of the Hidden Markov Modeling approach.

3.5 Constructing the chronology

Having obtained all the relevant probability distributions resulting from a run of the Forward-Backward algorithm, a next question arises on how to use all of this information to achieve an optimal chronology. Such a chronology should include both an optimal layer (i.e. age) at each depth as well as an uncertainty estimate of this layer number. There is no final answer to how these should be selected. A variety of reasonable choices can be made. In this section, I will describe what has been chosen here.

By way of the probabilities $\gamma_t(j)$, an annual layer probability distribution exist at each index t. These probability distributions include the entire information obtained, but to contemplate the development in these throughout an extended depth interval is not tractable. It is more convenient to summarize all the probability distributions by some descriptive statistics, and see how these evolve with depth in the ice core data.

To obtain an optimal layer at a given depth, which can provide an age estimate for the chronology, three choices come to mind: The mean, the median and the mode of the annual layer distributions. Using the mean is only a good approximation if the annual layer distributions are symmetric. If the distributions are skewed, the median or mode is a better choice. Thus, as there is no reason for the obtained layer distributions to be symmetric, the choice stood between the median and the mode of the distributions. Here, the mode was selected. The mode of the distribution will always be an integer, hence removing the need for working with years in the resulting chronology having non-integer values.

An estimate of the uncertainties was made by considering quantiles of the annual layer distributions, as these are better at describing skewed distributions than e.g. the standard deviation. The 25% and 75% quantiles (Q_{25} and Q_{75}) were used to provide a 50% confidence interval of the obtained age estimate, and the 2.5% and 97.5% quantiles ($Q_{2.5}$ and $Q_{97.5}$) produced a 95% confidence interval. These were used as descriptive statistics for the uncertainty of the age estimates.

Theoretically, the above definitions allow for the best estimate of the annual layer count at a certain depth (the mode) to be outside the confidence interval. However, for the data in consideration, this is not an issue. The mode and the median of the distributions are usually almost identical.

The uncertainty estimates produced by the Forward-Backward algorithm are computed based on the assumption that the annual layer detection algorithm provides an unbiased counting. Provided this is the case, an extra layer at one location is very likely to be counterbalanced by a layer lacking somewhere else. This means that although the uncertainty of the annual layer number continually increases, it will increase slower and slower with distance from the starting depth of the algorithm.

In contrast, the uncertainties in the manually counted GICC05 chronology were derived based on the counting of uncertain layers as ½±½ year. This uncertainty estimate does not assume an unbiased counting, and leads to an almost linear increase in uncertainty with depth. The best estimate of the involved uncertainties is most likely in between. The linear increase in uncertainty with depth is believed to be a very conservative estimate [*Andersen*

et al., 2006b], whereas the very narrow uncertainty interval band resulting from assuming an unbiased counting procedure most likely is very optimistic when dealing with real data.

3.6 The Viterbi algorithm

As mentioned previously, the definition of a "best" sequence of states is debatable. The Forward-Backward algorithm computes the sequence of states, in which each state individually maximizes the likelihood of the observed data – implying that the resulting state sequence may not even be allowed by the underlying model structure. In the layer detection model, it may e.g. be that one observation is most likely to be part of layer 2, whereas both surrounding observations are most likely to be part of layer 1. Such a state sequence contradicts the model assumption of layers to be in successive order.

Another definition of a "most likely state sequence" is used in the Viterbi algorithm [*Viterbi*, 1967], which provides the state sequence corresponding to the most likely segmentation of observations into annual layers. As such, this is a more meaningful definition given that it ensures the resulting state sequence to be valid. However, the Viterbi algorithm only computes the most likely layer boundaries, and does not keep track of the involved uncertainties. Consequently, it cannot be used to obtain an estimate of the uncertainties involved in the resulting timescale.

Furthermore, the most likely segmentation of a data series into annual layers does not necessarily imply the best counting of the number of annual layers in the data. In general, when applied to the visual stratigraphy data from NGRIP, the Viterbi algorithm tends to count fewer layers than the Forward-Backward algorithm. This is probably due to the existence of many possible layers which are not likely enough to be counted as layers by the Viterbi algorithm. In the Forward-Backward algorithm, on the other hand, all of these layers with low probability are slowly being summed up, and eventually an extra year is added.

Where distinct annual layers in a data series are visible, the annual layer estimates resulting from employing respectively the Forward-Backward algorithm and the Viterbi algorithm are very similar. However, in case of an ill-posed annual layer model or a data series containing a high degree of noise and many ambiguous layers, the two countings may differ significantly. By considering the results of the Viterbi algorithm as well as the Forward-Backward algorithm, the uncertainties involved in constructing the timescales can therefore be assessed.

The Forward-Backward algorithm generally ought to provide a better estimate of the timescale than the Viterbi algorithm. For the current purpose, most of the conclusions have therefore been based on the results of the Forward-Backward algorithm, and the main use of the results of the Viterbi algorithm has been as an indication of how well the chronology has been inferred. If, on the other hand, it is desired to divide up the observation sequence into years in order to obtain information on e.g. the seasonal signal in the data series, the Viterbi algorithm is the most promising.

The basic structure of the Viterbi algorithm is very similar to that of the Forward-Backward algorithm, and is also based on recursive evaluation of probabilities. The idea behind the algorithm is that the most likely state at a given time only depends on the most likely state before, and the transition probability between the previous and present state.

The most instructive way to visualize the procedure of the Viterbi algorithm is probably by use of a lattice (figure 3.6.1). Each possible state (or each generalized state in case of HSMMs) corresponds to a specific level in the lattice, and t increases towards the right. The Viterbi path is then the path through the lattice having the highest total probability when arriving to the end of the observation sequence. At each instant t, any path through the lattice arriving at a given state (ℓ_j, d) can be divided up into two parts: The path followed for reaching the previous (generalized) state, and the last path segment from the previous to the present state. The most probable partial path through the lattice ending exactly this state is the one, for which the product of the probabilities corresponding to each of these two path segments is the largest.

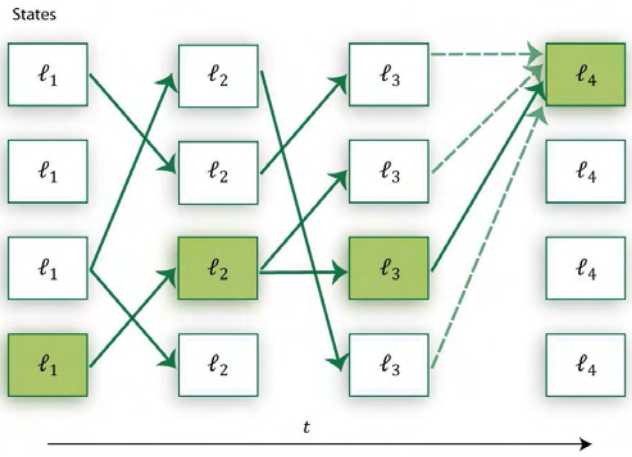


Figure 3.6.1: *A lattice of states. At each time step is shown the best partial path leading to each possible state (dark green arrows). For any given time t and state ℓ_j, the best partial path is the one which maximizes the following product: The probability of reaching a given previous state (ℓ_i) via the most likely path, times the probability of transitioning from this state (ℓ_i) to the considered present state (ℓ_j). By recursive computation, the most likely final state can be found. Using the information stored in the back-pointer, it is possible subsequently to go backwards in the state lattice and retrieve the most likely state sequence (colored light green).*

3.6.1 Partial path probabilities

We define $\delta_t(j, d)$ as the joint probability of the most probable state sequence in which state (ℓ_j, d) ends at t, and this first part of the observation sequence:

$$\begin{aligned}&\delta_t(j,d)\\&\equiv P\left(most\ probable\ state\ sequence\ s_{1:t}\ assuming\ that\ state\ \left(\ell_j,d\right)\ ends\ at\ t,\boldsymbol{o}_{1:t}\right)\\&=\max_{s_{1:t-d}} P\left(s_{1:t-d},S_{[t-d+1:t]}=\ell_j,\boldsymbol{o}_{1:t}\right)\end{aligned}$$

In the general case, this variable can be recursively calculated for $1 \leq t \leq T$ as (see box 3) [*Yu*, 2010]:

$$\delta_t(j,d) = \max_{\ell_i \in \mathcal{L} \setminus \ell_j,\, d' \in \mathcal{D}} \left\{ \delta_{t-d}(i,d')\; a_{(i,d')(j,d)}\; b_j(\boldsymbol{o}_{t-d+1:t}) \right\}$$

Note the similarities between this expression and that of the recursive estimation of the forward variable (3.4.4). The only difference is that the summation over all possible values of ℓ_i and d' in the computation of the forward variable $\alpha_t(j,d)$ here is being replaced with the maximum.

Using the definitions of $a_{(i,d')(j,d)}$ and $b_j(\boldsymbol{o}_{t-d+1:t})$ relevant to the annual layer detection algorithm, we arrive at the following simplified expression for partial path probabilities $\delta_t(j,d)$:

$$\begin{aligned} \delta_t(j,d) &= \max_{\ell_i \in \mathcal{L},\, d' \in \mathcal{D}} \left\{ \delta_{t-d}(i,d')\; a_{ij} p(d)\; b(\boldsymbol{o}_{t-d+1:t}) \right\} \\ &= \max_{d' \in \mathcal{D}} \left\{ \delta_{t-d}(j-1,d')\; p(d)\; b(\boldsymbol{o}_{t-d+1:t}) \right\} \\ &= \check{\delta}_{t-d}(j-1)\; p(d)\; b(\boldsymbol{o}_{t-d+1:t}) \end{aligned}$$

In the last equality, the notation $\check{\delta}_t(j) = \max_{d' \in \mathcal{D}} \{\delta_t(j,d')\}$ has been used. $\check{\delta}_t(j)$ can be interpreted as the joint probability of the most likely state sequence ending layer ℓ_j at time t and the corresponding first part of the observation sequence, i.e.:

$$\check{\delta}_t(j) \equiv \max_{s_{1:t-1}} P\left(s_{1:t-1}, S_{t]} = \ell_j, \boldsymbol{o}_{1:t}\right) = \max_{d' \in \mathcal{D}} \{\delta_t(j,d')\}$$

Initialization of the recursion, i.e. $\delta_{t \leq 0}(j,d)$, is similar to that of the forward variable (3.4.7).

3.6.2 Back-pointer

The variables $\delta_t(j,d)$ only estimates the resulting maximum probability of all partial paths ending state (ℓ_j, d) at t. In order to keep track of the state sequence giving rise to this maximum probability path, such information is simultaneously stored in a back-tracking vector variable, $\psi_t(j,d)$, with components $(\ell^\star, d^\star)$. In the general case, this backtracking vector can be computed by:

$$\psi_t(j,d) = (\ell^\star, d^\star) = \operatorname*{argmax}_{\ell_i \in \mathcal{L} \setminus \ell_j,\, d' \in \mathcal{D}} \left\{ \delta_{t-d}(i,d')\; a_{(i,d')(j,d)}\; b_j(\boldsymbol{o}_{t-d+1:t}) \right\}$$

In other words, whereas $\delta_t(j,d)$ tracks the resulting maximum partial path probabilities, the back-pointer keeps track of the previous (generalized) state which gave rise to these probabilities.

In the general case, this back-pointer includes information on the most likely duration $(d^\star)$ as well as state $(\ell^\star)$ corresponding to the previous generalized state. However, due to the fixed structure of the annual layer detection model, only information on the most likely duration of the previous layer is necessary for our application. In our case, the above simplifies to:

$$\psi_t(j,d) = (d^\star) = \arg\max_{d' \in \mathcal{D}} \{\delta_{t-d}(j-1,d')\}$$

Observing that t and d in the equation above only appears in the combination $t-d$, we can reduce the computational complexity by re-formulating the above using one less variable, namely:

$$\breve{\psi}_t(j) = (d^*) = \psi_{t+d}(j+1, d) = \arg\max_{d' \in \mathcal{D}} \{\delta_t(j, d')\}$$

For our case, this is a convenient notation. The original version of the back-pointer, $\psi_t(j,d)$, provides information on the most probable duration of the previous state $\ell_i = \ell_{j-1}$, under the assumption that state (ℓ_j, d) is ending at t. In contrast, the new back-pointer, $\breve{\psi}_t(j)$, provides information on the most likely duration of the present layer ℓ_j, when assuming it to end at t.

3.6.3 The back-tracking procedure

Having calculated these two parameters, the most likely path through the lattice of generalized states can be determined. First, the most likely final state is determined based on $\delta_t(j,d)$ (or, if possible, $\breve{\delta}_t(j)$). Subsequently, the $\breve{\psi}_t$'s are used repeatedly to obtain the most likely duration of this layer, thereby eventually determining the overall most likely state sequence.

First consider the simple case for which it is given that the last layer in the observation sequence ends exactly at T. The initialization of the backtracking procedure would then be the state ℓ_1^*, chosen as the state ending at T having the highest total probability:

$$\ell_1^* = \underset{\ell_j \in \mathcal{L}}{\operatorname{argmax}}\, \breve{\delta}_T(j) = \underset{\ell_j \in \mathcal{L}}{\operatorname{argmax}} \left(\max_{s_{1:t-1}} P\left(s_{1:t-1},\, S_{T]} = \ell_j, \boldsymbol{o}_{1:T}\right) \right)$$

Observe that the probabilities $\breve{\delta}_T(j)$ are based on the entire observation sequence. In the following, a star $(*)$ signifies the resulting most likely state, being indexed according to the number of steps taken backwards from T.

However, for the annual layer detection model applied on real data, no knowledge on the termination of the last layer is given. The last layer may terminate at any point at or after the end of the observation sequence. Hence, to find the most likely terminal state, we must take into account all options of ending time as well as annual layer number and duration. The initialization condition is therefore:

$$\begin{aligned}(t_1^*, \ell_1^*, d_1^*) = &\underset{\substack{t \geq T \\ \ell_j \in \mathcal{L} \\ d \geq t-T+1,\ d \in \mathcal{D}}}{\operatorname{argmax}} \quad \delta_t(j,d) \\ = &\underset{\substack{t \geq T \\ \ell_j \in \mathcal{L} \\ d \geq t-T+1,\ d \in \mathcal{D}}}{\operatorname{argmax}} \quad \left\{ \max_{(\ell)_{1:t-d}} P\left((\ell)_{1:t-d}, S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:T}\right) \right\}\end{aligned}$$

With t_1^* being the most likely ending time of the most likely state (ℓ_1^*, d_1^*).

As $\breve{\psi}_t(j)$ provides the most likely duration of the present layer ℓ_j when assuming this layer to end at t, the state sequence can now be traced back until reaching the start of the observation sequence. This is done by repeated application of:

$$(t_{k+1}^*, \ell_{k+1}^*, d_{k+1}^*) = \left(t_k^* - d_k^*,\ \ell_k^* - 1, \breve{\psi}_{t_{k+1}^*}(\ell_{k+1}^*)\right)$$

And the most likely segmentation of a data series into annual layers is hereby obtained.

Similarly to the Forward-Backward algorithm, also the Viterbi algorithm allows for a measure of the likelihood of the applied model and model parameters to be computed. The value of $\delta_{t_1^*}(j, d_1^*)$ gives the resulting joint probability of the most likely state sequence and the entire observation sequence. By maximizing this probability measure, the model and model parameters can be adjusted to best fit the observed data. Hence, also the Viterbi algorithm can be run in an unsupervised learning mode.

3.7 Implementation issues

In this section, some practical aspects on the implementation of the layer detection algorithms will be discussed. Firstly, it will be described how the layer detection algorithm can be treated to allow for sections of missing data. This turns out to be very easy. Secondly, it will be described how the computation of the probability variables in the layer detection algorithm very quickly is causing underflow due to machine precision, and how this issue was dealt with. And thirdly, I will mention the issue of how long it takes for the layer detection algorithm to compute the layering in a section of data.

3.7.1 Sections of missing data

Most data series are not complete, but contain areas of bad or missing data. For the visual stratigraphy, around 1% of the data series is lacking due to breaks in the ice core (figure 2.3.1). Fortunately, this does not present a major problem for the layer detection algorithm. The algorithm treats these sections in a way that much resembles what is usually done by eye in manual layer counting: The annual layers are interpolated based on information in the surrounding data and knowledge of the layer thickness distribution.

Data enters the layer detection algorithm through the probabilities $b(\boldsymbol{o}_{t+1:t+d})$:

$$b(\boldsymbol{o}_{t+1:t+d}) \equiv P\left(\boldsymbol{o}_{t+1:t+d} \mid S_{[t+1:t+d]} = \ell_j\right)$$

The simplest and crudest approximation (which is the one used here) is to assume no knowledge on the observations in areas of missing data, i.e. these data can have any value. Hence, when judging the likelihood of an annual layer based on a data segment containing missing data, the result only depends on how well the remaining data in the segment resemble a layer. If none of the data points exist, the likelihood is set equal to 1.

By doing so, the algorithm will fit in an appropriate number of layers in these sections, with the 'appropriate number' being determined based on knowledge of the layer thickness distribution in combination with the positions of annual layer boundaries in the surrounding data. Meanwhile, the uncertainty on the annual layer count may be increased. Small sections of missing data are of almost no importance for the resulting outcome of the algorithm, while for larger sections (much larger than the average layer thickness) the errors are increased along with the uncertainty estimates. This is very similar to what is done by eye when manually counting annual layers: Based on the surrounding layer

thicknesses, an appropriate number of years are added, and the estimated counting uncertainty is increased.

3.7.2 Preventing underflow

Given the definition of $\alpha_t(j,d)$ and β_t in the Forward-Backward algorithm as the joint probability of a specific layer and an increasingly long sequence of observations, a direct implementation of the forward and backward equations ((3.4.5) and (3.4.8)) will quickly suffer from underflow.

In order not to be limited by machine precision, one of two approaches are frequently used: The one most commonly encountered is to calculate the forward variable multiplied with a scaling function, and the backward variable multiplied with its inverse [*Yu*, 2010]. The dependence on the scaling function is then eliminated when the two variables are multiplied in the end. Being the fastest and most precise, this approach is often preferable. However, if there are relatively few state transitions compared to the length of the observation sequence, it may be difficult, perhaps even impossible, to find a suitable scaling function [*Yu*, 2010]. This is the case here. The annual layers are relatively thick and encompass many observations, and the use of a scaling function turned out to be inadequate. Hence, to solve this issue for the layer detection algorithm, a slightly different path than the customary one had to be taken. It turned out that a more practical approach was to use a log-transform of the forward-backward variables.

Inserting logs in the equations for $\alpha_t(j,d)$ and β_t gives:

$$\log \alpha_t(j,d) = \log p(d) + \log b(\boldsymbol{o}_{t-d+1:t}) + \log \tilde{\alpha}_{t-d}(j-1)$$

$$\log \tilde{\alpha}_{t-d}(j-1) = \log \sum_{d \in \mathcal{D}} \exp\left(\log \alpha_t(j,d)\right)$$

And for the backwards variable:

$$\log \beta_t = \log \sum_{d \in \mathcal{D}} \exp(\log p(d) + \log b(\boldsymbol{o}_{t-d+1:t}) + \log \beta_{t+d})$$

However, these equations cannot be directly implemented either. The recursive approach of summing up previously calculated values causes both of these equations to contain the operator $\log \sum \exp$, the required accuracy of which quickly exceeds machine precision. To prevent underflow, these values are therefore calculated using the following transformation:

$$\log \sum_i \exp(x_i) = \max(\boldsymbol{x}) + \log \sum_i \exp(x_i - \max(\boldsymbol{x}))$$

With $\boldsymbol{x}$ being the vector of all possible values x_i.

Computer-wise, this is an approximate solution only, as some of the very small terms may still drop out during the evaluation. However, the transformation ensures the largest and therefore most important terms to be included in the evaluation of the result. And for our purposes, the provided accuracy is more than sufficient.

Having accomplished the computation of $\alpha_t(j,d)$ and β_t without serious issues of underflow, it may still happen that the equations for calculating $\eta_t(j,d)$ and $\gamma_t(j,d)$ are subjected to underflow, simply due to very low probabilities of observing the exact observation sequence. All of these have therefore been evaluated in log-space.

To prevent underflow caused by limited machine precision, also the Viterbi equations have been implemented in log-space. In this case, however, the log-transformation is straight-forward: Given that the logarithm is a well-behaved and continuously increasing function, the logarithm of the maximal probabilities equals the maximum of the log-probabilities.

3.7.3 Execution time

A precarious spot of many Bayesian analyses is that they are very time consuming. Even though the Viterbi and the Forward-Backward algorithms provide an efficient procedure for computing all the relevant probabilities, the time issue still persists.

The computation of probabilities for a single batch of data containing 500 observations (with data in 1 mm resolution, as will be used later, this corresponds to 50 cm of data), does not take much time. In the present implementation of the algorithm, and for the conditions appropriate to the visual stratigraphy data later investigated, it only takes about 10-15 seconds on a laptop to simultaneously compute the layering as inferred by the Forward-Backward and the Viterbi algorithm.

However, one must bear in mind that if wanting to establish a chronology for several hundred meters of ice core, this will take a while. This is in particular the case if it is desired to take advantage of the ability of the algorithm to learn about the appropriate model parameters from the appearance of the data itself. In this case, the algorithm may be run in an iterative mode, as it will be described in chapter 5, which easily may lead to a 10-fold increase in computation time.

The computational burden lies in the calculation of the probabilities $b(\boldsymbol{o}_{t+1:t+d})$ (3.3.1), which evaluates whether a data segment is likely to be an annual layer or not. With an added complexity of the annual layer model, the addition of more data series etc., the computation time of these probabilities may increase. On the other hand, the computation of these probabilities has the potential to be parallelized, which may save some computation time.

To conclude: Most practical issues concerning the implementation of the layer detection algorithm have been solved. Missing data can be included in a simple yet efficient way, and by using log-probabilities, there is no serious problems with underflow. The only remaining issue is that it can be rather computationally demanding to use the algorithm on long sections of data. However, with the ever-expanding computer power available, this may not stay an issue for long.

4. Modeling the annual layers

This chapter deals with the fundamental problem posed in the HMM-based annual layer detection model: How can an annual layer be described in a simple way that allows us to evaluate the likelihood that a given observation segment is an annual layer? Knowing such probabilities for all possible start and ending locations is a prerequisite for inferring the most probable layering of an observation sequence. The most likely layer at any given location can then be inferred using the Forward-Backward algorithm (section 3.4), and the most likely segmentation of the observations into annual layers can be deduced from the Viterbi path procedure (section 3.6).

Obviously, the inferred layer boundaries depend critically on the applied annual layer model and model parameters (θ). The set of model parameters includes parameters used for judging how well a segment of observations $\boldsymbol{o}_{t_1:t_2}$ fits the characteristics of an annual layer ($b(\boldsymbol{o}_{t_1:t_2})$), as well as parameters describing the probability distribution of annual layer thicknesses ($p(d)$).

4.1 Annual layer thicknesses

Empirical data show that for a given depth interval, the annual layer thicknesses in an ice core, λ, are approximately lognormal distributed [*Andersen et al.*, 2006a]. The assumed probability distribution of the layer durations is therefore taken to be a lognormal distribution described by the two parameters μ_d and σ_d:

$$\lambda \sim p(d) = \mathrm{Log}\,\mathcal{N}(\mu_d, \sigma_d^2)$$

Throughout the following, log will denote the natural logarithm. To avoid confusion, the above two parameters will in the following be termed respectively the location parameter (μ_d) and the scale parameter (σ_d) of the distribution. An illustration of the effect of these on the layer thickness distribution is found in figure 4.1.1.

This continuous probability density function is discretized and normalized to provide duration probabilities corresponding to an integer number of data points. Furthermore,

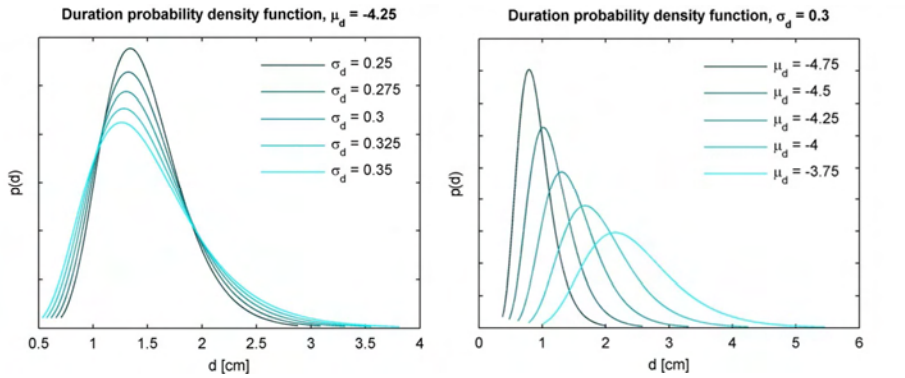


Figure 4.1.1: *A lognormal layer thickness probability distribution with different location (μ_d) and scale (σ_d) parameters.*

tails of the probability distribution are removed, such that there is a minimum as well as maximum duration allowed for each layer.

In *Andersen et al.* [2006b], the parameters governing the annual layer thickness distribution in the NGRIP ice core during different time periods were estimated (table 4.1.1). The flow-induced thinning of annual layers with depth has been taken into account by using strain-corrected mean values of $\log(\lambda)$. Consistent with previous studies [*D Dahl-Jensen et al.*, 1993], these strain-corrected values show the occurrence of significantly lower accumulation rates during the cold periods.

The variance of the distribution of $\log\lambda$ remains more or less the same throughout the three investigated time periods, indicating that the mechanisms of accumulation at the NGRIP site must be stable over time. The uniform variance of the distribution of strain-corrected layer thicknesses also applies to annual layer thicknesses at any given depth: As the applied strain correction is a multiplicative constant, the variance of the lognormal distribution of annual layer thicknesses is independent of this correction factor (see appendix A2). Hence, also the annual layer thickness distribution during the investigated periods was found to have a standard deviation around 0.3, more or less independently of depth and climate regime.

	GS-2	**Stadials**	**Interstadials**
$\boldsymbol{\mu_{corr}}$	-2.786	-2.899	-2.289
$\boldsymbol{\sigma_d}$	0.290	0.302	0.272

Table 4.1.1: *Parameter values governing the log-normal layer thickness distribution during different time periods and climate regimes. The value of μ_d at a given depth is dependent on the thinning of annual layers due to ice flow, which here has been corrected for by using a strain-corrected measure of the annual layer thicknesses (μ_{corr}). The corrected values provide an estimate of the mean of the original layer thickness distribution at time of deposition. In appendix A2 is included a derivation of the above quantities based on those given in Andersen et al. [2006b].*

4.2 The annual layer signature

The very core of the annual layer detection algorithm is the model for the annual layer signal. An appropriate choice for this is central to the accuracy of the resulting chronology. The annual signal model is used for calculating the emission probabilities:

$$b(\boldsymbol{o}_{t_1:t_2}) = P(\boldsymbol{o}_{t_1:t_2} | S_{[t_1:t_2]} = \ell_j, \theta), \;\; \ell_j \in \mathcal{L}$$

These provide an estimate of the likelihood that a given data segment represents an annual layer. The probability contribution due to the annual layer thickness itself is disregarded as it is taken care of separately by $p(d)$. While being among the most important parts of the algorithm, it is also the modeling of annual layers that presents the main shortcomings of the annual layer detection model described here. The remaining part of the algorithm is based on consistent mathematical principles, whereas this part dealing with actual data is inherently vague and difficult to properly define.

The annual layer model must be able to take into account the large degree of inter-annual variability in annual layer signal, while also being sufficiently simple. Simplicity is required given that the calculation of b takes up the major part of the computation time: b is to be calculated for all possible combinations of start and ending position of a layer within the entire observation sequence.

For the main part, the model equations derived in the previous sections present a simplification of the principal equations in Hidden Semi-Markov Modeling. This is not the case for the calculation of b, for which is needed a more sophisticated model than what is commonly used for Hidden Markov Model applications. In speech recognition, it is customary to use numerous observation sequences at once (all cepstral coefficients as well as their derivatives), but they are all assumed to follow relatively simple trajectories throughout each state. Here, we are faced with an annual layer signal which is relatively complex, and which cannot be modeled just by a constant value or a straight line.

4.2.1 An annual layer template

Each horizon in the visual stratigraphy record is assumed to be formed by gradual changes in dust influx to the ice sheet. With the dust influx displaying a seasonal variation, an obvious choice is to describe an annual layer by a smoothly changing mean value, which solely depends on the time of year of deposition. To simplify matters, linear transformation between time of year and depth within each layer is assumed, thereby eliminating the need to distinguish between the two.

The annual layers are modeled based on a generalized layer template, which consists of a selection of appropriate functions providing the general shape of the layer signal. The layer trajectories are then formed by linear combinations of these. However, the weighting corresponding to each of these functions is not fixed, but assumed to be Gaussian distributed around a given mean value. In this way, the template allows for a selected range of year to year variability in layer shapes. The employed model is an extended version of the linear trajectory models used by *Gish and Ng* [1996] and *Russell and Holmes* [1997], and bears many similarities to the one used for ECG waveform detection by *Kim et al.* [2004] and *Kim and Smyth* [2006].

The trajectory of an annual layer signal is described as the output of a linear system: A linear combination of basis functions, which themselves may be non-linear. These basis functions are combined to form a design matrix, X, which provides the generalized layer template. A one-year observation segment $\boldsymbol{O}_j$ (corresponding to a layer ℓ_j spanning d observations) is then modeled as:

$$\boldsymbol{O}_j \equiv \boldsymbol{o}_{t-d+1:t} = \mathrm{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \tag{4.2.1}$$

In the general case of a K-parameter annual layer model, $\boldsymbol{\beta}_j$ is the $K \times 1$ waveform parameter vector, and $\boldsymbol{\varepsilon}_j$ is a $d \times 1$ vector of residuals. The residuals will be assumed to be independent and identically distributed (i.i.d.) with zero-mean Gaussian distributions, i.e. $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_d(\mathbf{0}, \sigma_\varepsilon^2 \mathrm{I}_d)$, hence adding a Gaussian white noise component to the resulting observation vector $\boldsymbol{O}_j$. The size of the design matrix, X, is $d \times K$. The basis functions in the design matrix are normalized according to the annual layer thickness (d), such that the resulting shape of the annual layer signal is independent of the layer thickness.

Several options will be used as basis functions for the layer template to be contained in the design matrix X. The perhaps simplest (meaningful) basis is to assume the annual layer curve to be sinusoidal. However, it is a much better approximation to take into account possible differences in peak heights etc. by allowing a cosine to be overlain with a linear function with mean value 0 and a constant. Such assumption gives rise to the following design matrix:

$$\mathrm{X} = \begin{bmatrix} \cos(2\pi z_1) & z_1 - \frac{1}{2} & 1 \\ \cos(2\pi z_2) & z_2 - \frac{1}{2} & 1 \\ \cos(2\pi z_3) & z_3 - \frac{1}{2} & 1 \\ \vdots & \vdots & \vdots \\ \cos(2\pi z_d) & z_d - \frac{1}{2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{1}{2} & 1 \\ \cos(2\pi z_2) & z_2 - \frac{1}{2} & 1 \\ \cos(2\pi z_3) & z_3 - \frac{1}{2} & 1 \\ \vdots & \vdots & \vdots \\ 1 & \frac{1}{2} & 1 \end{bmatrix} \tag{4.2.2}$$

with $\boldsymbol{z} = (z_1, z_2, z_3, \ldots z_d)^\intercal = \left(0, \frac{1}{d-1}, \frac{2}{d-1}, \ldots, 1\right)^\intercal$ denoting the layer fractions corresponding to each of the d data points. $\intercal$ denotes the matrix transpose. The above is a three-parameter annual layer model, and correspondingly, the size of X is $d \times 3$.

Observe that this model by itself does not impose any restrictions regarding continuity of fitted trajectories across layer boundaries. For each layer, the most likely layer trajectory is chosen separately from that of the surrounding layer trajectories. Only if all basis functions have a value of zero in both ends, hereby forcing each layer trajectory also to have zero value here, the resulting fitted trajectories for successive layers will be continuous. However, even layer models which do not impose continuity across layer boundaries, turned out to work rather well in practice.

4.2.2 Allowing for inter-annual variability in layer shape

The annual layers in the visual stratigraphy data display a significant amount of variability in shape from one year to the next, and the annual signal model must be able to capture this diversity.

The approach taken here has been to divide up the layer shape characteristics into two levels: The uppermost level contains information on the average layer signal, and the lower one describes the individual differences in layer trajectory from this average layer shape. Such differences are called 'random effects'. In this way, it is possible to not only include information on how an average layer looks like, but also how much each individual layer is allowed to differ from this average signal. This kind of model is called a two-level hierarchical model. A conceptual illustration of such a model is found in figure 4.2.1.

In a hierarchical model, variations between individual annual layer signals are therefore caused by two different processes: One process is responsible for the general difference between individual layer trajectories due to different realizations of the shape form, while another process is responsible for the corruption by additive white noise on this layer shape. In other words: Two layer trajectories produced by the same set of overall model parameters need not be similar. But even if they are, detailed small-scale disparities in the realization of their trajectories will exist.

Here, the diversity of the individual layers has been accounted for by using a Bayesian approach. In Bayesian terms, the general idea of a hierarchical model can be formulated as a model which allow each layer to have its own parameters (lower level), but where these are coupled together by an overall population prior (upper level). In this way, only a generalized waveform template as the one in (4.2.2) is specified via the design matrix. The waveform parameter $\boldsymbol{\beta}_j$ itself is given as probability distributions.

Restricting ourselves to consider only multivariate Gaussian probability distribution as prior for the parameter values[5], this corresponds to assuming $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\varphi}, \Phi)$ in (4.2.1), where $\boldsymbol{\varphi}$ and Φ are two new parameters (replacing the one parameter $\boldsymbol{\beta}_j$). The likelihood that a segment comprises exactly one annual layer can then be evaluated by Bayesian linear regression (described in section 4.2.3). Bayesian linear regression will in general be superior to the ordinary least squares approach, which is prone to overfitting [*Bishop*, 2006].

To simplify the following derivations, this Bayesian annual layer model is now re-written as a two-level hierarchical linear model. This is done by splitting up the wave ters $\boldsymbol{\beta}_j$ into one part which describes the average layer signal ($\boldsymbol{\varphi}$), and a second part describing the random effect specific to each layer ($\boldsymbol{r}_j \in \mathcal{R}$):

$$\boldsymbol{\beta}_j = \boldsymbol{\varphi} + \boldsymbol{r}_j$$

[5] Such multivariate normal distribution being self-conjugate.

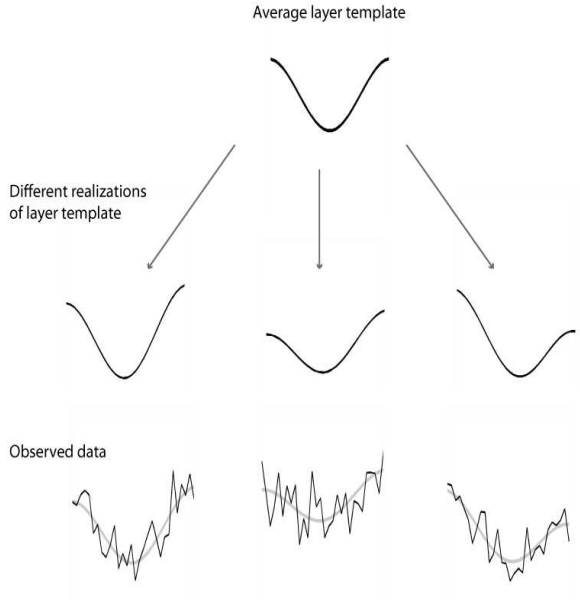


Figure 4.2.1: *A schematic drawing of the two-level hierarchical model used for modeling the annual layers. At the upper level, an average template for the layers is supplied. But based on this template, each layer has its own trajectory (middle level). Furthermore, white noise is added to the observed data (lowest level).*

With $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\varphi}, \Phi)$, the observation segment corresponding to an annual layer can be described as the output of the following linear system:

$$\begin{aligned} \boldsymbol{O}_j &= \mathrm{X}\boldsymbol{\beta}_j = \mathrm{X}\left(\boldsymbol{\varphi} + \boldsymbol{r}_j\right) + \boldsymbol{\varepsilon}_j \\ \boldsymbol{r}_j &\sim \mathcal{N}_K(\mathbf{0}, \Phi) \\ \boldsymbol{\varepsilon}_j &\sim \mathcal{N}_d(\mathbf{0}, \sigma_\varepsilon^2 \mathrm{I}_d) \end{aligned} \qquad (4.2.3)$$

This description is similar to the one used by *Kim and Smyth* [2006]. As the mean average layer signal is given by the parameter $\boldsymbol{\varphi}$, the mean of the random effect vector $\boldsymbol{r}_j$ is equal to zero. Its covariance is denoted Φ. The noise on each data point is assumed to be Gaussian white noise with variance σ_ε^2.

Hence, with this modification, the parameters contained in the parameter vector θ for the layer detection algorithm are the annual layer signal parameters $\boldsymbol{\varphi}$, Φ and σ_ε^2, along with the parameters describing the annual layer thickness distribution, μ_d and σ_d, i.e.: $\theta = \{\mu_d, \sigma_d, \boldsymbol{\varphi}, \Phi, \sigma_\varepsilon^2\}$.

4.2.3 Probability of a hypothesized annual layer segment

In order to evaluate the likelihood of an observation segment to be an annual layer, consider first an observation vector $\boldsymbol{O}_j$ of length d, which is known to be distributed according to a multivariate normal distribution with mean $\boldsymbol{\mu}$ (vector of length d) and covariance matrix Σ ($d \times d$ matrix):

$$\boldsymbol{O}_j \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$$

The probability density corresponding to obtaining an observation vector $\boldsymbol{O}_j$ is then given by [*Bishop*, 2006, p. 78]:

$$p(\boldsymbol{O}_j|\boldsymbol{\mu},\Sigma,d) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{O}_j-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{O}_j-\boldsymbol{\mu})\right) \tag{4.2.4}$$

Here, $|\Sigma|$ denotes the determinant of the matrix Σ, and Σ^{-1} is the matrix inverse. This is the general equation for the probability density function of a multivariate Gaussian. The hereby calculated values are probability densities, and they may therefore have values above 1.

If the value of $\boldsymbol{\beta}_j$ was known, such probabilities could be computed by inserting $\boldsymbol{\mu} = \mathrm{X}\boldsymbol{\beta}_j$, and $\Sigma = \sigma_\varepsilon^2\mathrm{I}_d$. However, as $\boldsymbol{\beta}_j$ is not known, we must first find the appropriate expressions for $\boldsymbol{\mu}$ and Σ for the linear annual layer model outlined in the previous section (4.2.3).

As the expectation value for $\boldsymbol{r}_j$ and $\boldsymbol{\varepsilon}_j$ both equal zero, $\mathbb{E}[\boldsymbol{r}_j] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\varepsilon}_j] = \mathbf{0}$, and $\boldsymbol{\varphi}$ is a constant, i.e. $\mathbb{E}[\boldsymbol{\varphi}] = \boldsymbol{\varphi}$**,** the expectation value for $\boldsymbol{O}_j$ is:

$$\mathbb{E}[\boldsymbol{O}_j] = \mathbb{E}[\mathrm{X}\boldsymbol{\varphi} + \mathrm{X}\boldsymbol{r}_j + \boldsymbol{\varepsilon}_j] = \mathrm{X}\boldsymbol{\varphi}$$

The corresponding covariance of $\boldsymbol{O}_j$ is given by:

$$\begin{aligned}
\mathrm{cov}[\boldsymbol{O}_j] &= \mathbb{E}\left[\left(\boldsymbol{O}_j - \mathbb{E}(\boldsymbol{O}_j)\right)\left(\boldsymbol{O}_j - \mathbb{E}(\boldsymbol{O}_j)\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\mathrm{X}\boldsymbol{r}_j + \boldsymbol{\varepsilon}_j\right)\left(\mathrm{X}\boldsymbol{r}_j + \boldsymbol{\varepsilon}_j\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\mathrm{X}\boldsymbol{r}_j\boldsymbol{r}_j\mathrm{X}^{\mathrm{T}} + \boldsymbol{\varepsilon}_j\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} + \mathrm{X}\boldsymbol{r}_j\boldsymbol{\varepsilon}_j^{\mathrm{T}} + \boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^{\mathrm{T}}\right] \\
&= \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j\boldsymbol{r}_j{}^{\mathrm{T}}\right]\mathrm{X}^{\mathrm{T}} + \mathbb{E}\left[\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^{\mathrm{T}}\right] \\
&= \mathrm{X}\Phi\mathrm{X}^{\mathrm{T}} + \sigma_\varepsilon^2\mathrm{I}_d
\end{aligned}$$

Here, the equality $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \Sigma$ [*Bishop*, 2006] has been utilized. As a consequence, the covariance matrix of any zero-mean vector $\boldsymbol{x}$ can be calculated as the expected value of the vector itself times its transpose.

Given that a conjugate prior was picked for the distribution of the wave parameters, also the distribution of the observation vector $\boldsymbol{O}_j$ will be Gaussian, and as shown above it has expectation value $\mathrm{X}\boldsymbol{\varphi}$ and covariance $\mathrm{X}\Phi\mathrm{X}^{\mathrm{T}} + \sigma_\varepsilon^2\mathrm{I}_d$. Consequently, the observation vector is described by the following probability distribution:

$$\boxed{\boldsymbol{O}_j \sim \mathcal{N}(\mathrm{X}\boldsymbol{\varphi}, \mathrm{X}\Phi\mathrm{X}^{\mathrm{T}} + \sigma_\varepsilon^2 I_d)}$$

With this probability distribution determined, the probability of observing a hypothesized layer can be calculated according to (4.2.4).

Note how the uncertainty in the precise value of $\boldsymbol{\beta}_j$ for a given layer acts to increase the allowed discrepancies from the mean trajectory, and it does so in a non-uniform way. The non-uniformity reflects how the linear regression ties in some sections of the trajectory more than others [*Bishop*, 2006].

The annual model described here is relatively simple. Yet, it is able to address the general problem faced by any annual layer detection model: The layers display a high degree of

variability. By the explicit modeling of this variability, it becomes an integrated part of the layer detection algorithm, and makes it able to better handle observation sequences with irregular and noisy annual layer signals.

4.3 Including the derivative of data series

In the previous section, an annual layer model was described, in which the layer signal was modeled as the noisy output of a linear system. However, the additive white noise assumption in this annual layer model is not very realistic. The white noise component of the model will be used to explain all variability in the observed data which cannot be explained by the chosen annual layer trajectory. As a consequence, the 'noise' on two consecutive observations will most likely be highly correlated, unlike additive white noise which per definition is uncorrelated. A more realistic model would have allowed the unexplained variability in the data series to be correlated. However, assuming correlated noise would make the calculations of $b(\boldsymbol{o}_{t_1:t_2})$ much more cumbersome, and therefore an assumption of white noise was preferred.

Fortunately, there is a relatively simple way to get around the unrealistic assumption of uncorrelated white noise. If considering instead the slope of the observation sequence, the noise here will be more like white noise than it was on the observation sequence itself. At the same time, however, the signal-to-noise ratio of the data will decrease, rendering the annual layer signal generally less perceptible. If considering the curvature of the data series, this would be even more so. The issue of correlated noise on the observations can therefore to some extent be addressed by simultaneously considering the observation sequence and its slope, and perhaps even its curvature, and locating the best annual layer boundaries based on the combined information in all of these. By simultaneously modeling the observation sequence and its derivatives, the impact of the white noise assumption is reduced, and there is no need to model the error correlation of successive observations. Thus, even though the information contained in the derivative of the observed data series also was present in the data series itself, information is added to the model by using this as an additional input. Indeed, when applied to real data, much better estimates of the annual layering were obtained when considering also the derivatives of the observed data.

Fortunately, it is easy to extend the previously derived equations to include information from more than a single data series. Consider e.g. the visual stratigraphy intensity data along with its derivatives. All of these contain a seasonal signal. Using the notation $\boldsymbol{y}$ for the original data vector, and $\Delta\boldsymbol{y}$ and $\Delta^2\boldsymbol{y}$ for respectively its slope and curvature, the observation vector $\boldsymbol{O}_j$ now consists of all three of these.

Extending the example from section 4.2.1, in which the annual layer signal was taken as proportional to a cosine function, the annual layer model can now be written in the following way:

$$\boldsymbol{O}_j = \begin{bmatrix} \boldsymbol{y} \\ \Delta\boldsymbol{y} \\ \Delta^2\boldsymbol{y} \end{bmatrix} = \mathrm{X}\boldsymbol{\beta}_j + \begin{bmatrix} \boldsymbol{\varepsilon}_j^{(1)} \\ \boldsymbol{\varepsilon}_j^{(2)} \\ \boldsymbol{\varepsilon}_j^{(3)} \end{bmatrix} = \mathrm{X}(\boldsymbol{\varphi} + \boldsymbol{r}_j) + \boldsymbol{E}_j$$

Here, $\boldsymbol{E}_j$ is the vector of white noise components for all data series. The appended design matrix is found by appending the derivatives of the basis functions to the original one (4.2.2):

$$\mathrm{X} = \begin{bmatrix} \cos(2\pi\boldsymbol{z}) & \boldsymbol{z} - \frac{1}{2} & \mathbf{1} \\ -\frac{2\pi}{d-1}\sin(2\pi\boldsymbol{z}) & \frac{\mathbf{1}}{d-1} & \mathbf{0} \\ -\left(\frac{2\pi}{d-1}\right)^2\cos(2\pi\boldsymbol{z}) & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Where the notation $\boldsymbol{z} = (z_1, z_2, z_3, \dots z_d)^\mathsf{T} = \left(0, \frac{1}{d-1}, \frac{2}{d-1}, \dots, 1\right)^\mathsf{T}$ has been used. In this way, the annual layer signal is described by the same number of parameters (i.e. the vectors $\boldsymbol{\varphi}$ and $\boldsymbol{r}_j$ remain the same size) as if only a single data series were used. But given that all three data series now are assumed to be described by the same set of parameter values, these must now be determined as those which fit all of these the best.

The noise on all data series is assumed to be additive Gaussian white noise. But the individual data series may have different noise levels. Hence, in the general case of M different data series, the combined vector of white noise components for all data series, $\boldsymbol{E}_j$, is distributed according to the following multivariate normal distribution:

$$\boldsymbol{E}_j = \begin{bmatrix} \boldsymbol{\varepsilon}_j^{(1)} \\ \boldsymbol{\varepsilon}_j^{(2)} \\ \vdots \\ \boldsymbol{\varepsilon}_j^{(M)} \end{bmatrix} \sim \mathcal{N}_{Md}(\mathbf{0}, \sigma_\varepsilon^2 \mathrm{W})$$

With W being a $Md \times Md$ diagonal matrix having the structure:

$$\mathrm{W} = \mathrm{diag}\,(\mathbf{1}_d, w_2\mathbf{1}_d, w_3\mathbf{1}_d, \dots, w_M\mathbf{1}_d)$$

The notation $\mathbf{1}_d$ has here been used for an all-ones vector of length d, and the weights w_m, $m \in \{1, \dots M\}$, denote the white noise variance on data sequence m relative to that of the first data sequence.

When taking into account the different noise levels corresponding to the individual data series, it can be shown that the resulting probability distribution for the observations $\boldsymbol{O}_j$, belonging to a hypothesized annual layer segment j, in this case is given by:

$$\boxed{\boldsymbol{O}_j \sim \mathcal{N}(\mathrm{X}\boldsymbol{\varphi}, \mathrm{X}\Phi\mathrm{X}^\mathsf{T} + \sigma_\varepsilon^2 \mathrm{W})}$$

Based on this probability distribution, the probability of observing a hypothesized layer can also in this case be calculated according to (4.2.4).

In the annual layer model applied here, an annual layer is modeled as the outcome of a linear model with additive white noise. But in reality the assumption of white noise is not really appropriate. The inadequacy of such assumption can be circumvented by supplementing the observation sequence with the sequence(s) of its derivatives. By doing so, much more accurate estimates of the annual layering in a data sequence can be made. The layer detection algorithm has therefore here been implemented in a semi-multi-parameter mode, in that not only the visual stratigraphy data itself, but also the derivatives of the data sequence, are taken into account.

4.4 Adding additional observation sequences

There are no conceptual difficulties in adding more data series to the annual layer recognition algorithm. These data series do not necessarily have to be connected to the visual stratigraphy. They may as well be e.g. multiple chemistry data series from CFA measurements, which contain an annual layer signal. In this way, an annual layer detection model based on Hidden Markov Modeling is able to provide a true multi-parameter annual layer counting algorithm, which can take into account the annual variation in many different chemical species in the ice core data at once.

Practical difficulties may arise, however, as it may be required to take the covariance between the signals in the individual data series into account. Doing so will cause the complexity of the model to increase. However, such issues may be minimized by selecting the employed data series carefully. A sensible choice could e.g. be to merely use data series which express different aspects of the annual signal.

Another issue is the different timing of peaks in the individual data series, and that not all measurements are supplied with an accurate depth scale. A very accurate depth scale is essential given that the annual layer thickness in the deeper part of the ice core is just a few centimeters.

The above mentioned obstacles are, however, still of minor character relative to the potential gains of obtaining an automated method of annual layer counting which may be able to compete with manual counting.

4.5 Possible extensions of annual layer model

Provided that increased computation time is not a serious concern for the annual layer detection algorithm, the model for how to calculate the likelihood of a hypothesized annual layer segment can be as complex as desired. In this respect, the annual layer model outlined above only presents a fairly simple implementation.

An obvious starting place for improving the modeling of an annual layer would be to allow for small inter-annual variations in the transformation between time of year and annual layer fraction. As the seasonal precipitation of snow changes from one year to the next, this by itself will give rise to a changing time-to-depth conversion for each layer – even if the seasonal variation in dust influx remains exactly the same.

In other words, even within each year, the time-to-depth conversion should be allowed some flexibility. Such time-warping within each annual layer can be allowed in several ways. One option is to warp the timeline within each hypothesized layer in order to conform the observations to the given template in the best possible way. Such approach can be pursued using the methods of Dynamical Time Warping (DTW), which often has been employed for such purpose within the realm of speech recognition [*Rabiner*, 1989].

Another approach could be to nest a tiny HMM model into the overall annual layer detection model. For each hypothesized layer segment, this HMM model can be used to both find the optimal warping specific to the current layer, as well as the resulting probability of the observation sequence to form an annual layer. As long as such a nested model is not too complex, it will not necessarily increase the computation time excessively, as the number of data points within each observation segment will be small.

Yet another method, by which the assumption of a linear time-depth relationship within each layer can be relaxed, is to model each layer segment as the outcome of a dynamical linear system. The prediction errors of the dynamical system can then be used to evaluate the probabilities of a given data segment to form an annual layer [*Ostendorf et al.*, 1996]. Dynamical systems can be made to allow for a wide range of annual layer signals, and has the advantage of allowing a direct modeling of the physical processes involved in forming the annual layers visible in the ice core data. Additionally, dynamical linear system theory also allow for the autoregressive character of the visual stratigraphy signal to be better exploited.

For the visual stratigraphy, one could also speculate that perhaps it would be an idea to use an indexing t based on the individual visible horizons in the line-scan images instead of depth. If creating a new data series by extracting one single data point per horizon, a more stable annual layer signal might emerge. However, such approach would only work for the visual stratigraphy, which is the only ice core record in which strata corresponding to individual snow events can be recognized.

5. Improving on layer parameter estimates

A major benefit of a HMM annual layer detection model lies in its ability to utilize the observations themselves to adjust and optimize the parameter values used for determining the most likely annual layering. In other words, the model is able to improve on an initial guess of the appropriate model parameters based on how the data actually looks like. By doing so, the model is able to continuously adjust itself to temporal changes in how an annual layer is expressed in the ice core data. Such adjustment is important due to the extreme abruptness of some climatic events recorded in the ice cores [*Steffensen et al.*, 2008], influencing both the mean annual layer thickness as well as the annual layer signal recorded in the data.

5.1 The optimal model parameters

As part of the Forward-Backward and the Viterbi algorithm, the likelihood of the current set of HMM parameter values, $L(\theta|\boldsymbol{o}_{1:T})$, is computed. This measure can be used to train the model by maximizing the likelihood of the joint set of parameters. The Maximum Likelihood (ML) value of the parameters is denoted θ_{ML}:

$$\begin{aligned}\theta_{ML} &= \underset{\theta}{\operatorname{argmax}} \log L(\theta|\boldsymbol{o}_{1:T}) \\ &= \underset{\theta}{\operatorname{argmax}} \log P(\boldsymbol{o}_{1:T}|\theta)\end{aligned}$$

The logarithm of the likelihood function is often used for convenience of easier calculations. As the logarithmic function is monotonously increasing, it bears no importance whether the likelihood or the log-likelihood is maximized.

The resulting annual layer parameters are those most likely to have produced a data sequence as the one observed. Yet, although not explicitly annotated above, the conditioning on the applied annual layer model should be kept in mind. The significance of the maximum likelihood parameters is contingent on the annual layer model to provide a reasonable description of an annual layer in the data. Also, no magic is involved. If for

some reason the annual layering is indistinguishable, the most likely annual layer parameters will be of little or no value. Likewise, even with an annual signal present and an appropriate annual layer model to detect it, the observation sequence may simply be too short to uncover a good estimate of the annual layer parameters.

Under challenging conditions, the performance of the annual layer detection algorithm may improve if supplied with any prior knowledge on the parameter values that we may be in possession of. Prior information on the parameter values can be taken into account by maximizing the posterior probability rather than the likelihood of the parameters. The Maximum a Posteriori (MAP) estimate of parameters is a point estimate corresponding to the mode of their posterior distribution:

$$\begin{aligned}\theta_{MAP} &= \underset{\theta}{\operatorname{argmax}}\ \log P(\theta|\boldsymbol{o}_{1:T}) \\ &= \underset{\theta}{\operatorname{argmax}}\ \log\big(P(\boldsymbol{o}_{1:T}|\theta)\ P(\theta)\big) \\ &= \underset{\theta}{\operatorname{argmax}}\ (\log P(\boldsymbol{o}_{1:T}|\theta) + \log P(\theta))\end{aligned}$$

Our prior knowledge on the value of the model parameters is contained within the probabilities $P(\theta)$.

Both Maximum Likelihood and Maximum a Posteriori optimization of the annual layer parameters can be achieved using a range of maximization methods. For Hidden Markov Modeling, the method most commonly used is the Expectation-Maximization (EM) algorithm, which within the framework of HMMs is also known as the Baum-Welch algorithm [*Leonard E. Baum et al.*, 1970; *Dempster et al.*, 1977; *Gupta and Chen*, 2011; *Welch*, 2003].

The EM-algorithm attempts to find a point estimate of the most likely set of parameter values. Hence, it is not a Bayesian optimization method, by which the entire posterior probability distribution would be determined. However, a straight-forward adaptation of the EM-algorithm makes it possible to include a prior for the parameter values, and the algorithm can thus be implemented in a semi-Bayesian way. It hereby has obvious advantages compared to e.g. optimization using the Newton-Raphson method [*Press*, 1996], while being much faster than a full Bayesian Markov-Chain Monte Carlo optimization procedure [*Mosegaard and Tarantola*, 1995; *Tarantola*, 2005].

5.2 The Expectation-Maximization algorithm

The basic idea behind the Expectation-Maximization algorithm is as follows: By comparing a first evaluation of the hidden state sequence with observations, a new estimate of the model parameters can be obtained. Given that this set of parameters is influenced by the observations, these will provide a better assessment of the true parameter values than the initial guess. By repeating this exercise multiple times, an optimal set of parameters can be found.

The EM-algorithm (figure 5.2.1) hence alternates between two steps: First step is the expectation step (E-step), in which the current set of model parameters, $\theta^{(k)}$, is used for

calculating the conditional expectation of the log-likelihood of the joint set of parameter values:

$$Q(\theta|\theta^{(k)}) \equiv \mathbb{E}\big[\log L\big(\theta|\mathfrak{D}_{complete}\big) \mid \boldsymbol{o}_{1:T}, \theta^{(k)}\big]$$

Here, $\mathfrak{D}_{complete}$ denotes the complete data set, which consists of the observation sequence as well as all hidden sequences, and the expectation value is taken with respect to all of the hidden sequences. In case only the state sequence, $s_{1:T}$, is hidden, we have:

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= \mathbb{E}\big[\log L(\theta|s_{1:T}, \boldsymbol{o}_{1:T}) \mid \boldsymbol{o}_{1:T}, \theta^{(k)}\big] \\ &= \sum_{s_{1:T}\in\mathcal{L}^T} P\big(s_{1:T}\big|\boldsymbol{o}_{1:T}, \theta^{(k)}\big)\ \log L(\theta|s_{1:T}, \boldsymbol{o}_{1:T}) \end{aligned} \tag{5.2.1}$$

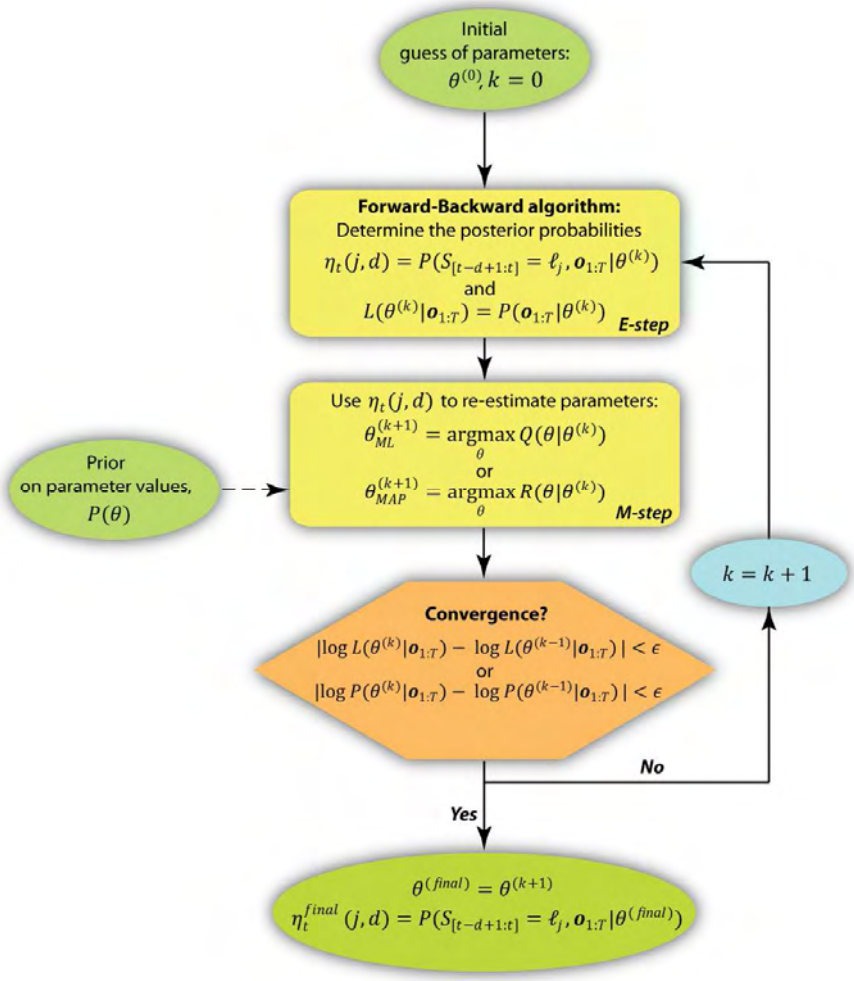


Figure 5.2.1: *Flow chart depicting the procedure of the EM-algorithm in the general case of the Forward-Backward algorithm. Starting from an initial guess of the model parameters, $\theta^{(0)}$, the joint set of these is continually being improved upon by iterating between the E-step and the M-step. During the E-step, posterior probabilities based on the current set of model parameters are computed, and using these, a new and better set of parameters is estimated during the M-step by maximizing either their resulting likelihood or posterior probability. The two steps are repeated until convergence, at which stage a (local) maximum of the likelihood/posterior probability function has been reached.*

Although intimidating as this definition of $Q(\theta|\theta^{(k)})$ may appear, it is only a repetition of what was said in words above: From the observations – pretending that our current guess of model parameters, $\theta^{(k)}$, is correct – the probability of any hidden state sequence $s_{1:T}$ can be computed using the Forward-Backward algorithm. Assume for the sake of simplicity that only a single hidden state sequence is likely to occur. (Likewise, this could e.g. be the output of the Viterbi algorithm, which only determines a single optimal state sequence). By comparing this state sequence with the observations, the log-likelihood of any model parameter value can then be evaluated. This is $Q(\theta|\theta^{(k)})$. When multiple hidden state sequences are conceivable, the expectation of the log-likelihood is calculated in order to take into account the different probabilities associated with the respective hidden state sequences. Given that the Viterbi algorithm does not provide any such probability estimates of alternative state sequences, its performance in this regard is inferior. For that reason, the EM-algorithm will here only be used in combination with the Forward-Backward algorithm.

As implied by its notation, $Q(\theta|\theta^{(k)})$ is a function of the model parameters θ. But it also depends implicitly on the current guess of the joint set of model parameters $\theta^{(k)}$ used for estimating the hidden state sequence in the first place.

Secondly, after having calculated the Q-function in the E-step, a maximization step (M-step) is performed. During the M-step, the Q-function is used for selecting an improved set of model parameters, $\theta^{(k+1)}$. If a maximum likelihood estimate is desired, the new set of model parameters is chosen as the joint set maximizing the Q-function:

$$\theta_{ML}^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(k)})$$

Alternatively, if a Maximum a Posteriori estimate of the parameters is required, the maximization step is achieved by maximizing the auxiliary function $R(\theta|\theta^{(k)})$ defined by [*Gauvain et al.*, 1994]:

$$R\left(\theta\middle|\theta^{(k)}\right) = Q\left(\theta\middle|\theta^{(k)}\right) + \log P(\theta)$$

$$\theta_{MAP}^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, R(\theta|\theta^{(k)}) \qquad (5.2.2)$$

These two steps may now be iterated. The convergence properties of the EM-algorithm are analogous for both types of estimates.

It can be theoretically proven (see derivation in appendix A3) that by iterating between the E- and M-steps, the likelihood of the parameter values will never decrease. And hopefully, the parameters will simultaneously converge towards their optimal values. But notice the word 'hopefully'. *Gupta and Chen* [2011, p.227] give the following explanation of why the function to be optimized has been termed the Q-function: *"We like to say that the Q stands for quixotic because it is a bit crazy and hopeful and beautiful to think you can find the maximum likelihood estimate of θ in this way that iterates round-and-round like a windmill, and if Don Quixote had been a statistician, it is just the sort of thing he might have done"*.

No guarantee is given that the EM-algorithm will manage to locate the globally most likely set of parameter values. The deterministic behavior of the algorithm may e.g. cause it to get trapped in a local maximum of the likelihood function. Furthermore, not even convergence of the sequence $\{\theta^{(k)}\}$ is guaranteed. The EM-algorithm only assures the convergence of the sequence $\{\log L(\theta^{(k)}|\boldsymbol{o}_{1:T})\}$ provided that this sequence is bounded [*Gupta and Chen*, 2011]. A detailed discussion on the convergence issues of the EM-algorithm can be found in *Wu* [1983].

Despite these theoretical limitations, numerous implementations of the EM-algorithm have demonstrated that in practice the EM-algorithm often does a good job (see e.g. *Lagendijk et al.* [1990], *Snyder and Politte* [1983], *Zabin and Poor* [1991]). The troublesome ability of the algorithm to get caught up in a local maximum of the likelihood function can to some extent be addressed by using a couple of random starts, and ultimately picking the set of parameters having obtained the highest likelihood value.

For the annual layer detection model developed here, it will in practice make most sense to run the EM-algorithm in Maximum a Posteriori mode: Prior probabilities can be estimated based on previous data, and by adding the information contained in these, the robustness of the algorithm will in general increase. In addition, the usage of a prior allows the use of relatively short observation sequences, which may not contain sufficient information to produce robust statistics and therefore reliable Maximum Likelihood estimates. As short observation sequences can be processed faster, this may speed up the annual layer detection significantly.

When optimizing the model parameters used in the layer detection algorithm, the procedure therefore should be: First, an initial set of parameters is used as input to the Forward-Backward algorithm, and a proposed first segmentation of the observations into annual layers is obtained. This state sequence, along with the corresponding observations and our priors, is now used to compute the a posteriori most probable set of parameter values. Subsequently, this set is used as input for a new iteration of the Forward-Backward algorithm. The iteration continues until the sequence has converged. In this way, the Forward-Backward algorithm can be trained by the observations to obtain a 'best' estimate of the involved parameters.

In the next section, the Maximum Likelihood re-estimation equations of the annual layer parameters will be derived. Although related to similar equations for other applications of the EM-algorithm (see e.g. *Chien and Huang* [2003], *Kim and Smyth* [2006]), the equations here have been developed for the specific assumptions applicable for annual layer detection in ice cores (lognormal layer thicknesses etc.). Having obtained the Maximum Likelihood re-estimation equations, it is only a little step further in complexity to derive also the Maximum a Posteriori update equations (section 5.4).

5.3 Maximum Likelihood layer parameters

The simplest way to re-estimate the annual layer signal parameters is by considering the observations corresponding to each layer individually, these being defined using the

proposed most likely layer boundaries. Such an approach is sometimes termed a 'hard count'. However, the Forward-Backward algorithm does not as such calculate the most likely segmentation of the observations; what is estimated are the probabilities of each observation to belong to a given layer (se discussion in section 3.6). As a result, each layer boundary constitutes a probability distribution, and taking these probabilities into account (making a 'soft count') will lead to a better estimate of the parameter values.

A soft count approach ensures a proper treatment of sections within the observation sequence in which the annual layering cannot be determined very accurately, and accordingly neither can the corresponding annual layer parameters. In a hard count, a parameter estimate derived from such a section of the observation sequence would be a single, and most likely incorrect, value. Using a soft count, the result is a probability distribution of the full suite of potential parameter values, in which each annual layer parameter is weighted according to the probability of the corresponding segment to form an annual layer. In this way, an uncertainty in the layer boundary positions is transferred to the estimated parameter values.

In chapter 4, the annual layer signal model used for parameterizing the annual layers was described. In short, each layer is considered a noisy outcome of a generalized linear model with K base functions, and annual layer thicknesses are assumed log-normally distributed. According to this, the annual layer parameters are: Parameters describing the layer thickness distribution (μ_d, σ_d), parameter vector describing the mean annual layer signal ($\boldsymbol{\varphi}$) along with the covariance matrix hereof (Φ), and variance of the white noise component (σ_ε^2). Hence, the joint set of parameters, whose likelihood is to be estimated, is $\theta = \{\mu_d, \sigma_d, \boldsymbol{\varphi}, \Phi, \sigma_\varepsilon^2\}$.

According to the definition of the Q-function, parameters must be re-estimated using their expectations:

$$Q(\theta|\theta^{(k)}) \equiv \mathbb{E}\left[\log L\left(\theta|\mathfrak{D}_{complete}\right) \mid \boldsymbol{o}_{1:T}, \theta^{(k)}\right]$$

First of all, the complete dataset, $\mathfrak{D}_{complete}$, which contain the observation sequence as well as all hidden sequences, must be identified.

In (5.2.1) the state sequence $s_{1:T}$ was the hidden sequence. We will now bring into play instead the generalized state sequence $q_{1:N}$, where $q_n = (\ell_n, d_n)$, $\ell_n \in \mathcal{L}$, $d_n \in \mathcal{D}$. N is the total number of annual layers in the observation sequence. The observation segment belonging to the n'te layer will be denoted $\boldsymbol{O}_n$ (in case of several observation es, $\boldsymbol{O}_n$ contains the complete set of observations corresponding to that layer) and the entire observation sequence can therefore be written $\boldsymbol{O}_{1:N}$. Observe that this notation implies knowledge on the segmentation of the observation sequence into annual layers, which is not implied by the notation $\boldsymbol{o}_{1:T}$.

As the succession of layers is fixed, and all layers are modeled alike, the information contained in consecutive values of ℓ_n is trivial. All necessary information lies within the durations of the individual layers. The hidden sequence will therefore be taken as the sequence of layer durations, $d_{1:N} \in \mathcal{D}^N$, which gives a complete description of the segmentation of the observation sequence into annual layers.

Additionally, the random effect $\boldsymbol{r}_n \in \mathcal{R}$ corresponding to each layer will be considered a hidden variable. The random effect vector describes the differences in shape of the individual layers from the mean signal, and it has a Gaussian distribution, $\boldsymbol{r}_n \sim \mathcal{N}_K(\boldsymbol{0}, \Phi)$. The corresponding hidden sequence is $\boldsymbol{r}_{1:N} \in \mathcal{R}^N$. In our case, the complete data is therefore: $\mathfrak{D}_{complete} = \{d_{1:N}, \boldsymbol{r}_{1:N}, \boldsymbol{O}_{1:N}\}$.

The log-likelihood of a joint set of parameters θ, when conditioned on the complete data, can be calculated as:

$$\begin{aligned}
&\log L\big(\theta|\mathfrak{D}_{complete}\big) = \log L(\theta|d_{1:N}, \boldsymbol{r}_{1:N}, \boldsymbol{O}_{1:N}) \\
&= \log P(d_{1:N}, \boldsymbol{r}_{1:N}, \boldsymbol{O}_{1:N}|\, \mu_d, \sigma_d, \boldsymbol{\varphi}, \Phi, \sigma_\varepsilon^2) \\
&= \log\big(P(d_{1:N}|\, \mu_d, \sigma_d, \boldsymbol{\varphi}, \Phi, \sigma_\varepsilon^2)\, P(\boldsymbol{r}_{1:N}|d_{1:N}, \mu_d, \sigma_d, \boldsymbol{\varphi}, \Phi, \sigma_\varepsilon^2) \\
&\qquad\qquad \cdot P(\boldsymbol{O}_{1:N}|d_{1:N}, r_{1:N},\, \mu_d, \sigma_d, \boldsymbol{\varphi}, \Phi, \sigma_\varepsilon^2)\big) \\
&= \log P(d_{1:N}|\mu_d, \sigma_d) + \log P(\boldsymbol{r}_{1:N}|\Phi) + \log P(\boldsymbol{O}_{1:N}|d_{1:N}, \boldsymbol{r}_{1:N}, \boldsymbol{\varphi}, \sigma_\varepsilon^2)
\end{aligned}$$

Consequently, the Q-function can be decomposed into three parts:

$$Q\big(\theta\big|\theta^{(k)}\big) = Q_1\big(\mu_d, \sigma_d\big|\theta^{(k)}\big) + Q_2\big(\Phi\big|\theta^{(k)}\big) + Q_3\big(\boldsymbol{\varphi}, \sigma_\varepsilon^2\big|\theta^{(k)}\big) \tag{5.3.1}$$

With:

$$\begin{aligned}
Q_1\big(\mu_d, \sigma_d\big|\theta^{(k)}\big) &\equiv \mathbb{E}\big[\log P(d_{1:N}|\mu_d, \sigma_d)\ |\ \boldsymbol{o}_{1:T}, \theta^{(k)}\big] \\
Q_2\big(\Phi\big|\theta^{(k)}\big) &\equiv \mathbb{E}\big[\log P(\boldsymbol{r}_{1:N}|\Phi)\ |\ \boldsymbol{o}_{1:T}, \theta^{(k)}\big] \\
Q_3\big(\boldsymbol{\varphi}, \sigma_\varepsilon^2\big|\theta^{(k)}\big) &\equiv \mathbb{E}\big[\log P(\boldsymbol{O}_{1:N}|d_{1:N}, \boldsymbol{r}_{1:N}, \boldsymbol{\varphi}, \sigma_\varepsilon^2)\ |\ \boldsymbol{o}_{1:T}, \theta^{(k)}\big]
\end{aligned} \tag{5.3.2}$$

Hereby, the log-likelihood has been decoupled into three different parts. As each of the five layer signal parameters in (5.3.1) is included in just a single term of $Q\big(\theta\big|\theta^{(k)}\big)$, the terms (5.3.2) can be maximized separately to get a re-estimate of the respective parameter values. Yet, their resulting optimum values depend on the entire set of previous parameters, $\theta^{(k)}$.

The expectation values must be calculated with respect to all possible realizations of both two hidden sequences, $d_{1:N}$ and $\boldsymbol{r}_{1:N}$. The probability of obtaining a specific realization is:

$$\begin{aligned}
P\big(d_{1:N}, \boldsymbol{r}_{1:N}|\boldsymbol{o}_{1:T}, \theta^{(k)}\big) &= P\big(d_{1:N}\big|\boldsymbol{o}_{1:T}, \theta^{(k)}\big)\, P\big(\boldsymbol{r}_{1:N}\big|d_{1:N}, \boldsymbol{o}_{1:T}, \theta^{(k)}\big) \\
&= P\big(d_{1:N}\big|\boldsymbol{o}_{1:T}, \theta^{(k)}\big)\, P\big(\boldsymbol{r}_{1:N}\big|\boldsymbol{O}_{1:N}, \theta^{(k)}\big) \\
&= P\big(d_{1:N}\big|\boldsymbol{o}_{1:T}, \theta^{(k)}\big) \prod_{n=1}^{N} P\big(\boldsymbol{r}_n\big|\boldsymbol{O}_n, \theta^{(k)}\big)
\end{aligned} \tag{5.3.3}$$

In the second step of the derivation above, the duration sequence $d_{1:N}$ was used for segmenting the observations into their respective layers: A conditioning on $\{d_{1:N}, \boldsymbol{o}_{1:T}\}$ is equal to a conditioning on $\boldsymbol{O}_{1:N}$, which includes information on the layer boundary positions. For the factorization in the last step, it was utilized that the individual values of $\boldsymbol{r}_n$ corresponding to each layer are assumed independent.

The layer durations are only allowed to take on discrete values, specified as the number of observations covered by each layer. The expectation value of the log-likelihood with respect to the hidden state sequence $d_{1:N}$ is therefore found by summing up the probability

contributions from each of these. On the other hand, the random components $\boldsymbol{r}_j$ are continuously valued. To take the expectation with respect to the hidden state sequence composed of these, the corresponding probability densities must be integrated.

With these prerequisites in hand, the re-estimation equations for each of the five annual layer signal parameters will now be evaluated.

5.3.1 Layer thickness parameters

To re-estimate the duration distribution parameters μ_d and σ_d, we must evaluate the conditional expectation of $\log P(d_{1:N}|\mu_d, \sigma_d)$, summed and integrated over all possible realizations of the hidden sequences. This expectation can be computed as follows:

$$
\begin{aligned}
&Q_1\big(\mu_d, \sigma_d \big| \theta^{(k)}\big) = \mathbb{E}\big[\log P(d_{1:N}|\mu_d, \sigma_d) \mid \boldsymbol{o}_{1:T}, \theta^{(k)}\big] \\
&= \sum_{d_{1:N} \in \mathcal{D}^N} \int_{\boldsymbol{r}_{1:N} \in \mathcal{R}^N} P\big(d_{1:N}, \boldsymbol{r}_{1:N} \big| \boldsymbol{o}_{1:T}, \theta^{(k)}\big)\ \log P(d_{1:N}|\mu_d, \sigma_d)\, d\boldsymbol{r}_{1:N} \\
&= \sum_{d_{1:N} \in \mathcal{D}^N} \int_{\boldsymbol{r}_{1:N} \in \mathcal{R}^N} \Bigg(P\big(d_{1:N} \big| \boldsymbol{o}_{1:T}, \theta^{(k)}\big) \cdot P\big(\boldsymbol{r}_{1:N} \big| \boldsymbol{O}_{1:N}, \theta^{(k)}\big) \\
&\qquad\qquad \cdot \sum_{n=1}^{N} \log P(d_n|\mu_d, \sigma_d) \Bigg) d\boldsymbol{r}_{1:N} \\
&= \sum_{n=1}^{N} \sum_{d_{1:N} \in \mathcal{D}^N} \int_{\boldsymbol{r}_{1:N} \in \mathcal{R}^N} P\big(d_{1:N} \big| \boldsymbol{o}_{1:T}, \theta^{(k)}\big)\, P\big(\boldsymbol{r}_{1:N} \big| \boldsymbol{O}_{1:N}, \theta^{(k)}\big) \log P(d_n|\mu_d, \sigma_d)\, d\boldsymbol{r}_{1:N}
\end{aligned}
$$

Consider first only the contribution from the n'te layer to the sum:

$$
\begin{aligned}
&\sum_{d_{1:N} \in \mathcal{D}^N} \int_{\boldsymbol{r}_{1:N} \in \mathcal{R}^N} P\big(d_1, d_2, \dots d_{n-1}, d_n, d_{n+1}, \dots d_N \big| \boldsymbol{o}_{1:T}, \theta^{(k)}\big)\, P\big(\boldsymbol{r}_{1:N} \big| \boldsymbol{O}_{1:N}, \theta^{(k)}\big)\ \log P(d_n|\mu_d, \sigma_d)\, d\boldsymbol{r}_{1:N} \\
&= \sum_{d_n \in \mathcal{D}} P\big(d_n \big| \boldsymbol{o}_{1:T}, \theta^{(k)}\big)\ \log P(d_n|\mu_d, \sigma_d)
\end{aligned}
$$

with d_n being the duration of layer n. The equality holds as all other variables ($\boldsymbol{r}_{1:N}$ and $d_{1:n-1}, d_{n+1:N}$) can be marginalized out, and hence do not contribute to the sum. As the above is true for all n layers, we see that:

$$
Q_1\big(\mu_d, \sigma_d \big| \theta^{(k)}\big) = \sum_{n=1}^{N} \sum_{d_n \in \mathcal{D}} P(d_n|\boldsymbol{o}_{1:T}, \theta^{(k)})\ \log P(d_n|\mu_d, \sigma_d) \tag{5.3.4}
$$

To evaluate this sum directly, knowledge on the duration probabilities of each individual layer (calculated based on the entire observation sequence and the current set of parameters) is required. However, such probabilities have not been determined. Instead, the probability of ending layer j with duration d at index t was computed using the Forward-Backward algorithm:

$$
\bar{\eta}_t(j, d) = P\big(S_{[t-d+1:t]} = \ell_j \big| \boldsymbol{o}_{1:T}, \theta^{(k)}\big)
$$

Seeking a way to utilize this knowledge, equation (5.3.4) can be transformed by including the probability of ending the n'te at t_n, and summing over all values of ending time t_n:

$$\begin{aligned} Q_1\left(\mu_d, \sigma_d \middle| \theta^{(k)}\right) &= \sum_{n=1}^{N} \sum_{d_n \in \mathcal{D}} P\left(d_n \middle| \boldsymbol{o}_{1:T}, \theta^{(k)}\right) \log P(d_n | \mu_d, \sigma_d) \\ &= \sum_{t_n=1}^{T} \sum_{n=1}^{N} \sum_{d_n \in \mathcal{D}} P\left(d_n, t_n \middle| \boldsymbol{o}_{1:T}, \theta^{(k)}\right) \log P(d_n | \mu_d, \sigma_d) \\ &= \sum_{t_n=1}^{T} \sum_{n=1}^{N} \sum_{d_n \in \mathcal{D}} P\left(S_{[t_n - d_n + 1:t_n]} = \ell_n \middle| \boldsymbol{o}_{1:T}, \theta^{(k)}\right) \log P(d_n | \mu_d, \sigma_d) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{d=1}^{D} \bar{\eta}_t(j, d) \log P(d | \mu_d, \sigma_d) \end{aligned} \tag{5.3.5}$$

Note that in the last equality, the summation over the different layers is indexed by j instead of n, with j going up to J (not N). There is a subtle difference between these two: We do not know the value of N, which is the actual number of annual layers in the observation sequence (if we did, there was no need for the analysis in the first place!). But a maximum number of annual layers in the sequence, J, can be estimated. Fortunately, there is no problem in summing over all J possible layers.

Finally, we are ready for the M-step of re-estimating the duration parameters (μ_d and σ_d) by optimizing the Q-function with respect to these. As previously mentioned, Q_1 is the only part of the Q-function depending on these parameters. Their optimum values can therefore be found by differentiating equation (5.3.5) with respect to each of the two duration parameters, and setting the result equal to zero.

The layer thickness parameter d is assumed to follow a (discretized) lognormal distribution described by the location parameters μ_d and scale parameter σ_d:

$$P(d | \mu_d, \sigma_d) \propto \frac{1}{d\sqrt{2\pi\sigma_d^2}} \exp\left(\frac{-(\log d - \mu_d)^2}{2\sigma_d^2}\right)$$

And thus:

$$\log P(d | \mu_d, \sigma_d) = -\log\left(d\sqrt{2\pi\sigma_d^2}\right) - \frac{(\log d - \mu_d)^2}{2\sigma_d^2} + \text{constant}$$

Differentiating the Q-function (5.3.1) with respect to μ_d gives:

$$\begin{aligned} \frac{\partial Q\left(\theta \middle| \theta^{(k)}\right)}{\partial \mu_d} &= \frac{\partial Q_1\left(\mu_d, \sigma_d \middle| \theta^{(k)}\right)}{\partial \mu_d} \\ &= \sum_{t,j,d} \bar{\eta}_t(j, d) \frac{\partial}{\partial \mu_d} \left(-\log\left(d\sqrt{2\pi\sigma_d^2}\right) - \frac{(\log d - \mu_d)^2}{2\sigma_d^2} + \text{constant}\right) \end{aligned}$$

$$= \frac{1}{\sigma_d^2} \sum_{t,j,d} \bar{\eta}_t(j,d)(\log d - \mu_d) \tag{5.3.6}$$

By setting this expression equal to zero, the value of μ_d which maximizes the Q-function (re-estimated parameter values will throughout the following be denoted with a '$\widehat{\ }$') is found to be:

(5.3.7)
$$\boxed{\hat{\mu}_d = \frac{\sum_{t,j,d} \bar{\eta}_t(j,d) \log d}{\sum_{t,j,d} \bar{\eta}_t(j,d)}}$$

Differentiating Q with respect to σ_d yields:

$$\begin{aligned}\frac{\partial Q\left(\theta \middle| \theta^{(k)}\right)}{\partial \sigma_d} &= \frac{\partial Q_1\left(\mu_d, \sigma_d \middle| \theta^{(k)}\right)}{\partial \sigma_d} \\ &= \sum_{t,j,d} \bar{\eta}_t(j,d) \frac{\partial}{\partial \sigma_d} \left(-\log\left(d\sqrt{2\pi\sigma_d^2} \right) - \frac{(\log d - \mu_d)^2}{2\sigma_d^2} + constant \right) \\ &= -\frac{1}{\sigma_d} \sum_{t,j,d} \bar{\eta}_t(j,d) \left(1 - \frac{(\log d - \mu_d)^2}{\sigma_d^2} \right)\end{aligned}$$

We hence arrive at the following rule for how the value of σ_d^2 should be re-estimated:

(5.3.8)
$$\boxed{\hat{\sigma}_d^2 = \frac{\sum_{t,j,d} \bar{\eta}_t(j,d)(\log d - \hat{\mu}_d)^2}{\sum_{t,j,d} \bar{\eta}_t(j,d)}}$$

Here, $\hat{\mu}_d$ is the new estimate (as given by (5.3.7)) of the location parameter governing the lognormal distribution of annual layer thicknesses.

Regardless of the trouble we went through to derive these re-estimation equations, (5.3.7) and (5.3.8), for the two duration parameters, both of these have a rather straightforward interpretation. The best estimate of μ_d (mean of the distribution of log-transformed layer thicknesses) is simply a weighted sample average of the logarithm to the layer durations, with each segment being weighted according to the probability of it to form an annual layer. The uncertainties in individual layer boundary positions are reflected in the probabilities $\bar{\eta}_t(j,d)$. Likewise, the best estimate of the scale parameter of the distribution (variance of the distribution of log-transformed layer thicknesses) is just the weighted sample variance. As one might have guessed beforehand, these are the values of the duration parameters with the highest likelihood based on the current segmentation of the data series into annual layers.

5.3.2 Covariance of the random effects

In analogy to the way that Q_1 was treated in the previous section, also the remaining parts of the Q-function can be re-arranged by summing over all possible realizations of the hidden sequences, and subsequently marginalizing out as many hidden variables as possible. For the second part of the Q-function, which deals with the covariance matrix of the random effect vector, we get:

$$
\begin{aligned}
&Q_2\left(\Phi\middle|\theta^{(k)}\right) = \mathbb{E}\left[\log P(\boldsymbol{r}_{1:N}|\Phi) \mid \boldsymbol{o}_{1:T}, \theta^{(k)}\right] \\
&= \sum_{d_{1:N}\in\mathcal{D}^N} \int_{\boldsymbol{r}_{1:N}\in\mathcal{R}^N} P\left(d_{1:N}, \boldsymbol{r}_{1:N}|\boldsymbol{o}_{1:T}, \theta^{(k)}\right) \log P(\boldsymbol{r}_{1:N}|\Phi)\, d\boldsymbol{r}_{1:N} \\
&= \sum_{d_{1:N}\in\mathcal{D}^N} \int_{\boldsymbol{r}_{1:N}\in\mathcal{R}^N} P\left(d_{1:N}\middle|\boldsymbol{o}_{1:T}, \theta^{(k)}\right) \left(\prod_{n=1}^{N} P\left(\boldsymbol{r}_n\middle|\boldsymbol{O}_n, \theta^{(k)}\right)\right) \sum_{n=1}^{N} \log P(\boldsymbol{r}_n|\Phi)\, d\boldsymbol{r}_{1:N} \\
&= \sum_{d_{1:N}\in\mathcal{D}^N} P\left(d_{1:N}\middle|\boldsymbol{o}_{1:T}, \theta^{(k)}\right) \sum_{n=1}^{N} \left(\int_{\boldsymbol{r}_{1:N}\in\mathcal{R}^N} \prod_{n=1}^{N} P\left(\boldsymbol{r}_n\middle|\boldsymbol{O}_n, \theta^{(k)}\right) \log P(\boldsymbol{r}_n|\Phi)\, d\boldsymbol{r}_{1:N}\right) \\
&= \sum_{n=1}^{N} \int_{\boldsymbol{r}_n\in\mathcal{R}} P\left(\boldsymbol{r}_n\middle|\boldsymbol{O}_n, \theta^{(k)}\right) \log P(\boldsymbol{r}_n|\Phi)\, d\boldsymbol{r}_n \\
&= \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{d=1}^{D} \int_{\boldsymbol{r}_n\in\mathcal{R}} P\left(S_{[t-d+1:t]} = \ell_n\middle|\boldsymbol{o}_{1:T}, \theta^{(k)}\right) P\left(\boldsymbol{r}_n\middle|\boldsymbol{O}_n, \theta^{(k)}\right) \log P(\boldsymbol{r}_n|\Phi) d\boldsymbol{r}_n \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{d=1}^{D} \int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\, P\left(\boldsymbol{r}_j\middle|\boldsymbol{O}_j, \theta^{(k)}\right) \log P\left(\boldsymbol{r}_j\middle|\Phi\right) d\boldsymbol{r}_j
\end{aligned}
$$

In the above, $\boldsymbol{r}_j$ is the random component corresponding to a proposed layer segment $\boldsymbol{O}_j$. It is a vector with K components, K being the number of base functions used for modeling the annual layer signal. The random components are assumed to be distributed according to a multivariate normal distribution with mean vector $\mathbf{0}$ and $K \times K$ covariance matrix Φ, i.e. $\boldsymbol{r}_j \sim \mathcal{N}_K(\mathbf{0}, \Phi)$. The corresponding probability density function is:

$$P\left(\boldsymbol{r}_j|\Phi\right) = (2\pi)^{-\frac{K}{2}} |\Phi|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}\boldsymbol{r}_j\right)$$

Hence, the log-probability density corresponding to a given value of $\boldsymbol{r}_j$ is:

$$\log P\left(\boldsymbol{r}_j|\Phi\right) = -\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\Phi| - \frac{1}{2}\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}\boldsymbol{r}_j$$

Inserting this expression into Q_2, and differentiating with respect to the matrix Φ, we get:

$$\frac{\partial Q\left(\theta\middle|\theta^{(k)}\right)}{\partial\Phi} = \frac{\partial Q_2\left(\Phi\middle|\theta^{(k)}\right)}{\partial\Phi}$$

$$
\begin{aligned}
&= \sum_{t,j,d} \int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\, P\left(\boldsymbol{r}_j\middle|\boldsymbol{O}_j, \theta^{(k)}\right) \frac{\partial}{\partial\Phi}\left(-\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\Phi| - \frac{1}{2}\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}\boldsymbol{r}_j\right) d\boldsymbol{r}_j \\
&= \sum_{t,j,d} \int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\, P\left(\boldsymbol{r}_j\middle|\boldsymbol{O}_j, \theta^{(k)}\right) \left(-\frac{1}{2}\frac{\partial\log|\Phi|}{\partial\Phi} - \frac{1}{2}\frac{\partial(\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}\boldsymbol{r}_j)}{\partial\Phi}\right) d\boldsymbol{r}_j
\end{aligned}
$$

Knowing that the covariance matrix Φ (and consequently also its inverse) is symmetric, we have the following two identities [*Petersen and Pedersen*, 2008]:

$$\frac{\partial\log|\Phi|}{\partial\Phi} = (\Phi^{-1})^{\mathrm{T}} = \Phi^{-1}$$

$$\frac{\partial(\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}\boldsymbol{r}_j)}{\partial\Phi} = -(\Phi^{-1})^{\mathrm{T}}\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}(\Phi^{-1})^{\mathrm{T}} = -\Phi^{-1}\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}$$

The insertion of these two identities leads to the following expression for the derivative of the Q-function with respect to the random effect covariance matrix Φ:

$$\begin{aligned}
\frac{\partial Q(\theta|\theta^{(k)})}{\partial\Phi} &= \sum_{t,j,d}\int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\, P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\left(-\frac{1}{2}\Phi^{-1}+\frac{1}{2}\Phi^{-1}\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\Phi^{-1}\right)d\boldsymbol{r}_j \\
&= -\frac{1}{2}\sum_{t,j,d}\bar{\eta}_t(j,d)\left(\int_{\boldsymbol{r}_j\in\mathcal{R}} P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\left(\mathrm{I}_K-\Phi^{-1}\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\right)d\boldsymbol{r}_j\right)\Phi^{-1} \\
&= -\frac{1}{2}\left(\sum_{t,j,d}\bar{\eta}_t(j,d)-\Phi^{-1}\sum_{t,j,d}\bar{\eta}_t(j,d)\int_{\boldsymbol{r}_j\in\mathcal{R}} P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\,\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}d\boldsymbol{r}_j\right)\Phi^{-1} \\
&= -\frac{1}{2}\left(\sum_{t,j,d}\bar{\eta}_t(j,d)-\Phi^{-1}\sum_{t,j,d}\bar{\eta}_t(j,d)\,\mathbb{E}\big[\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\big|\boldsymbol{O}_j,\theta^{(k)}\big]\right)\Phi^{-1}
\end{aligned}$$

Setting this result equal to zero (i.e. all entries in the resulting matrix must be zero), yields the following update equation for the random effect covariance matrix:

$$\begin{aligned}
\sum_{t,j,d}\bar{\eta}_t(j,d)-\hat{\Phi}^{-1}\sum_{t,j,d}\bar{\eta}_t(j,d)\,\mathbb{E}\big[\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}|\boldsymbol{O}_j,\theta^{(k)}\big] = 0 \;\Leftrightarrow \\
\hat{\Phi}\sum_{t,j,d}\bar{\eta}_t(j,d) = \sum_{t,j,d}\bar{\eta}_t(j,d)\,\mathbb{E}\big[\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\,|\boldsymbol{O}_j,\theta^{(k)}\big] \;\Leftrightarrow
\end{aligned}$$

$$\boxed{\hat{\Phi} = \frac{\sum_{t,j,d}\bar{\eta}_t(j,d)\,\mathbb{E}\big[\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\,|\boldsymbol{O}_j,\theta^{(k)}\big]}{\sum_{t,j,d}\bar{\eta}_t(j,d)}} \tag{5.3.9}$$

This is the covariance matrix with the maximum likelihood. Also this re-evaluated covariance matrix of the random component has a nice interpretation, although perhaps slightly less intuitive as for the layer thickness distribution parameters. The most likely covariance matrix is the sample average of the expected value of $\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}$ for each possible segment, weighted according to the probability of the segment to be an annual layer. For a normally distributed vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ, the expectation value of $\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ has the following interpretation:

$$\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+\Sigma,$$

Hence, as the vectors $\boldsymbol{r}_j$ on average are assumed to have zero mean, what is calculated is the averaged predicted covariance matrix of the random component for the proposed layer sequence.

5.3.3 Mean trajectory parameter and white noise component

Finally, also the third and last part of the Q-function can be maximized with respect to the two remaining variables: The mean trajectory parameter ($\boldsymbol{\varphi}$) and the variance of the white noise component (σ_ε^2). The basic idea behind uncovering their optimum values is similar

to the above, although the derivation itself is somewhat more involved. Once again, the part of the Q-function which involves these parameter values must first be expressed in terms of $\bar{\eta}_t(j,d)$:

$$
\begin{aligned}
&Q_3\big(\boldsymbol{\varphi},\sigma_\varepsilon^2\big|\theta^{(k)}\big) = \mathbb{E}\big[\log P(\boldsymbol{O}_{1:N}|d_{1:N},\boldsymbol{r}_{1:N},\boldsymbol{\varphi},\sigma_\varepsilon^2) \mid \boldsymbol{o}_{1:T},\theta^{(k)}\big] \\
&= \sum_{d_{1:N}\in\mathcal{D}^N}\int_{\boldsymbol{r}_{1:N}\in\mathcal{R}^N} P\big(d_{1:N},\boldsymbol{r}_{1:N}|\boldsymbol{o}_{1:T},\theta^{(k)}\big)\,\log P(\boldsymbol{O}_{1:N}|d_{1:N},\boldsymbol{r}_{1:N},\boldsymbol{\varphi},\sigma_\varepsilon^2)\,d\boldsymbol{r}_{1:N} \\
&= \sum_{d_{1:N}\in\mathcal{D}^N}\int_{\boldsymbol{r}_{1:N}\in\mathcal{R}^N} P\big(d_{1:N}\big|\boldsymbol{o}_{1:T},\theta^{(k)}\big)\prod_{n=1}^{N} P\big(\boldsymbol{r}_n\big|\boldsymbol{O}_n,\theta^{(k)}\big)\sum_{n=1}^{N}\log P(\boldsymbol{O}_n|d,\boldsymbol{r}_n,\boldsymbol{\varphi},\sigma_\varepsilon^2)\,d\boldsymbol{r}_{1:N} \\
&= \sum_{n=1}^{N}\int_{\boldsymbol{r}_n\in\mathcal{R}} P\big(\boldsymbol{r}_n\big|\boldsymbol{O}_n,\theta^{(k)}\big)\,\log P(\boldsymbol{O}_n|d,\boldsymbol{r}_n,\boldsymbol{\varphi},\sigma_\varepsilon^2)\,d\boldsymbol{r}_n \\
&= \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{d=1}^{D}\int_{\boldsymbol{r}_n\in\mathcal{R}} P\big(S_{[t-d+1:t]}=\ell_n\big|\boldsymbol{o}_{1:T},\theta^{(k)}\big)\,P\big(\boldsymbol{r}_n\big|\boldsymbol{O}_n,\theta^{(k)}\big)\log P(\boldsymbol{O}_n|d,\boldsymbol{r}_n,\boldsymbol{\varphi},\sigma_\varepsilon^2)d\boldsymbol{r}_n \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{d=1}^{D}\int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\,P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\log P\big(\boldsymbol{O}_j\big|d,\boldsymbol{r}_j,\boldsymbol{\varphi},\sigma_\varepsilon^2\big)\,d\boldsymbol{r}_j \qquad (5.3.10)
\end{aligned}
$$

As described in section 4.2, each annual layer is parameterized as the noisy outcome of a generalized linear model with mean parameter $\boldsymbol{\varphi}$ and random component $\boldsymbol{r}_j$:

$$\boldsymbol{O}_j = \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)+\boldsymbol{E}_j$$

where $\boldsymbol{O}_j$ is a vector containing all observations in the M data series which are part of layer ℓ_j, and the noise vector $\boldsymbol{E}_j$ corresponds to this assembled observation vector. Both of these are vectors of length Md. The noise on the data series is assumed to be Gaussian white noise, but the individual data series may have different noise levels. Hence, the noise vector is distributed according to a multivariate normal distribution $\boldsymbol{E}_j \sim \mathcal{N}_{Md}(\mathbf{0},\Sigma_E)$ where Σ_E is a $Md\times Md$ diagonal matrix with the structure:

$$\Sigma_E = \sigma_\varepsilon^2\mathrm{W} \quad \text{with} \quad \mathrm{W} = \mathrm{diag}\,(\mathbf{1}_d, w_2\mathbf{1}_d, w_3\mathbf{1}_d, \dots, w_M\mathbf{1}_d)$$

The matrix W is assumed known. In the above, the notation $\mathbf{1}_d$ has been used for an all-ones vector of length d.

If the random effect for a layer is given, the layer expression in the data series is only modified by the addition of white noise. In this case, the probability density corresponding to observing a segment of observations $\boldsymbol{O}_j$ covering exactly one layer is given by:

$$
\begin{aligned}
&P\big(\boldsymbol{O}_j|\boldsymbol{\varphi},\boldsymbol{r}_j,\sigma_\varepsilon^2,d\big) \\
&\qquad = (2\pi)^{-\frac{Md}{2}}\,|\Sigma_E|^{-\frac{1}{2}}\,\exp\left(-\frac{1}{2}\Big(\boldsymbol{O}_j-\mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)^{\mathsf{T}}\Sigma_E^{-1}\Big(\boldsymbol{O}_j-\mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)\right)
\end{aligned}
$$

In terms of the weight matrix W, the determinant and inverse of Σ_E can be written as:

$$|\Sigma_E| = |\sigma_\varepsilon^2\mathrm{W}| = \sigma_\varepsilon^{2Md}|\mathrm{W}|$$

$$\Sigma_E^{-1} = (\sigma_\varepsilon^2\mathrm{W})^{-1} = \sigma_\varepsilon^{-2}\mathrm{W}^{-1}$$

Inserting these expressions into the equation above and taking the log, the corresponding log-probability is found to be:

$$\log P\big(\boldsymbol{O}_j|\boldsymbol{\varphi},\boldsymbol{r}_j,\sigma_\varepsilon^2,d\big) = \\ -\frac{Md}{2}\log 2\pi - Md\log\sigma_\varepsilon - \frac{1}{2}\log|\mathrm{W}| \; -\frac{1}{2\sigma_\varepsilon^2}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)^{\mathsf{T}}\mathrm{W}^{-1}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)$$

Now, differentiating the Q-function (5.3.1) with respect to $\boldsymbol{\varphi}$, we get:

$$\begin{aligned}\frac{\partial Q\big(\theta\big|\theta^{(k)}\big)}{\partial\boldsymbol{\varphi}} &= \frac{\partial Q_3\big(\boldsymbol{\varphi},\sigma_\varepsilon^2\big|\theta^{(k)}\big)}{\partial\boldsymbol{\varphi}} \\ &= \sum_{t,j,d}\int_{\boldsymbol{r}_j\in\boldsymbol{\mathcal{R}}}\bar{\eta}_t(j,d)\;P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\;\frac{\partial}{\partial\boldsymbol{\varphi}}\big(\log P\big(\boldsymbol{O}_j|\boldsymbol{\varphi},\boldsymbol{r}_j,\sigma_\varepsilon^2,d\big)\big)\,d\boldsymbol{r}_j \\ &= -\frac{1}{2\sigma_\varepsilon^2}\sum_{t,j,d}\int_{\boldsymbol{r}_j\in\boldsymbol{\mathcal{R}}}\bar{\eta}_t(j,d)\;P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big) \\ &\quad\cdot\frac{\partial}{\partial\boldsymbol{\varphi}}\left(\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)^{\mathsf{T}}\mathrm{W}^{-1}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)\right)d\boldsymbol{r}_j\end{aligned}$$

By completing the square, and differentiating the individual terms separately, it can be seen that the following identity holds:

$$\frac{\partial}{\partial\boldsymbol{\varphi}}\left(\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)^{\mathsf{T}}\mathrm{W}^{-1}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)\right) = -2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\big)$$

A formal derivation of this result is included in appendix A4.1.

Inserting the above identity, it is seen that:

$$\begin{aligned}\frac{\partial Q\big(\theta\big|\theta^{(k)}\big)}{\partial\boldsymbol{\varphi}} &= \frac{1}{\sigma_\varepsilon^2}\sum_{t,j,d}\int_{\boldsymbol{r}_j}\bar{\eta}_t(j,d)\;\;P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\,\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\boldsymbol{r}_j\big)\Big)\,d\boldsymbol{r}_j \\ &= \frac{1}{\sigma_\varepsilon^2}\sum_{t,j,d}\bar{\eta}_t(j,d)\;\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi}+\int_{\boldsymbol{r}_j}P\big(\boldsymbol{r}_j\big|\boldsymbol{O}_j,\theta^{(k)}\big)\,\boldsymbol{r}_j d\boldsymbol{r}_j\right)\right) \\ &= \frac{1}{\sigma_\varepsilon^2}\sum_{t,j,d}\bar{\eta}_t(j,d)\;\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\boldsymbol{\varphi}+\mathbb{E}\big[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\big]\big)\Big) \qquad (5.3.11)\end{aligned}$$

The optimal re-estimated value for $\boldsymbol{\varphi}$, denoted $\widehat{\boldsymbol{\varphi}}$, is now found by setting this expression equal to zero:

$$\begin{aligned}&\sum_{t,j,d}\bar{\eta}_t(j,d)\;\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\Big(\boldsymbol{O}_j - \mathrm{X}\big(\widehat{\boldsymbol{\varphi}}+\mathbb{E}\big[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\big]\big)\Big) = \mathbf{0} \;\Leftrightarrow \\ &\sum_{t,j,d}\bar{\eta}_t(j,d)\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\widehat{\boldsymbol{\varphi}} = \sum_{t,j,d}\bar{\eta}_t(j,d)\;\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\big(\boldsymbol{O}_j - \mathrm{X}\mathbb{E}\big[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\big]\big) \;\Leftrightarrow\end{aligned}$$

$$\boxed{\widehat{\boldsymbol{\varphi}} = \left(\sum_{t,j,d} \bar{\eta}_t(j,d)\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\right)^{-1} \sum_{t,j,d} \bar{\eta}_t(j,d)\,\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)} \tag{5.3.12}$$

Neglecting the added complexity caused by the summing up of matrices, and just considering the contribution from a single proposed layer, we get:

$$\begin{aligned}\mathrm{X}\widehat{\boldsymbol{\varphi}} &= \mathrm{X}(\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\\ &= \mathrm{X}(\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\mathrm{X}^{\mathsf{T}}(\mathrm{X}\mathrm{X}^{\mathsf{T}})^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\\ &= \boldsymbol{O}_j - \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\end{aligned}$$

By comparison to the parameterization of an annual layer, it is seen that this is precisely the most likely value of the mean trajectory parameter $\boldsymbol{\varphi}$ for this proposed segment:

$$\mathrm{X}\widehat{\boldsymbol{\varphi}} = \boldsymbol{O}_j - \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right] \quad \Leftrightarrow \quad \boldsymbol{O}_j = \mathrm{X}\left(\widehat{\boldsymbol{\varphi}} + \mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$

Hence, also this update equation can be interpreted in a sensible way as the overall most likely value of the parameter vector $\boldsymbol{\varphi}$**,** when weighted with the annual layer boundary probabilities provided by the Forward-Backward algorithm.

A similar exercise can be done to retrieve the best estimate for the observed value of the white noise variance, σ_ε^2, based on the current segmentation of the data series. The derivative of the Q-function with respect to σ_ε is given by:

$$\begin{aligned}&\frac{\partial Q\left(\theta|\theta^{(k)}\right)}{\partial\sigma_\varepsilon} = \frac{\partial Q_3\left(\boldsymbol{\varphi},\sigma_\varepsilon^2|\theta^{(k)}\right)}{\partial\sigma_\varepsilon}\\ &= \sum_{t,j,d}\int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\,P\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\frac{\partial}{\partial\sigma_\varepsilon}\left(\log P\left(\boldsymbol{O}_j|\boldsymbol{\varphi},\boldsymbol{r}_j,\sigma_\varepsilon^2,d\right)\right)d\boldsymbol{r}_j\\ &= \sum_{t,j,d}\int_{\boldsymbol{r}_j\in\mathcal{R}} \bar{\eta}_t(j,d)\,P\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\cdot\frac{\partial}{\partial\sigma_\varepsilon}\left(-\frac{Md}{2}\log 2\pi - Md\log\sigma_\varepsilon - \frac{1}{2}\log|\mathrm{W}|\right.\\ &\qquad\left.-\frac{1}{2\sigma_\varepsilon^2}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)\right)^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)\right)\right)d\boldsymbol{r}_j\\ &= \frac{1}{\sigma_\varepsilon}\sum_{t,j,d}\bar{\eta}_t(j,d)\cdot\end{aligned}$$

$$\left(-Md + \frac{1}{\sigma_\varepsilon^2}\int_{\boldsymbol{r}_j\in\mathcal{R}} P\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)\right)^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)\right)d\boldsymbol{r}_j\right)$$

Using $\boldsymbol{E}_j = \boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)$, the above can be rewritten in a much more compact way, namely:

$$\frac{\partial Q\left(\theta|\theta^{(k)}\right)}{\partial\sigma_\varepsilon} = \frac{1}{\sigma_\varepsilon}\sum_{t,j,d}\bar{\eta}_t(j,d)\left(-Md + \frac{1}{\sigma_\varepsilon^2}\mathbb{E}\left[\boldsymbol{E}_j^{\mathsf{T}}\mathrm{W}^{-1}\boldsymbol{E}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$

Finding the optimal value of σ_ε^2 by setting this derivative of the Q-function equal to zero, leads to the following expression:

$$\sum_{t,j,d} \bar{\eta}_t(j,d) Md = \frac{1}{\hat{\sigma}_\varepsilon^2} \sum_{t,j,d} \bar{\eta}_t(j,d) \mathbb{E}\left[\boldsymbol{E}_j^\intercal \mathrm{W}^{-1} \boldsymbol{E}_j | \boldsymbol{O}_j, \theta^{(k)}\right] \Leftrightarrow$$

$$\boxed{\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t,j,d} \bar{\eta}_t(j,d) \mathbb{E}\left[\boldsymbol{E}_j^\intercal \mathrm{W}^{-1} \boldsymbol{E}_j | \boldsymbol{O}_j, \theta^{(k)}\right]}{\sum_{t,j,d} \bar{\eta}_t(j,d) Md}} \quad (5.3.13)$$

Again, this update equation can be interpreted in a reasonable way, as the variable $\boldsymbol{E}_j$ is a vector describing the estimated residuals from the parameterized layer trajectory. In case of just a single data series, i.e. $M = 1$, the matrix W (and hence also W^{-1}) is the identity matrix. Consider a single proposed layer. With all residuals described by the same zero-mean normal distribution, the best estimate of their variance can be calculated as the mean of squared residuals, i.e.:

$$\frac{1}{d} \sum_{i=1}^{d} \boldsymbol{E}_j(i)^2 = \frac{1}{d} \boldsymbol{E}_j^\intercal \boldsymbol{E}_j$$

where d is the layer duration, and $\boldsymbol{E}_j(i)$ is the i'te component of the vector $\boldsymbol{E}_j$. Based on the corresponding segmentation probabilities derived from the Forward-Backward algorithm, the above equation (5.3.13) then calculates a weighted average of these.

If using M data series which do not share the same white noise variance, their corresponding residuals must be evaluated according to their relative noise levels, as they are given in the W-matrix. As W is diagonal, the term $\boldsymbol{E}_j^\intercal \mathrm{W}^{-1} \boldsymbol{E}_j$ can be calculated by:

$$\boldsymbol{E}_j^\intercal \mathrm{W}^{-1} \boldsymbol{E}_j = \sum_{i=1}^{Md} \boldsymbol{E}_j(i)^2 / w(i)$$

with $w(i)$ being the i'te entry on the diagonal of W. The average of the weighted squared residuals are then found by dividing with Md, which is the total number of observations in the assembled observation vector $\boldsymbol{O}_j$. In this way, the W-matrix normalizes the residuals of the individual data series to the noise level of data series number one, and the overall best estimate for σ_ε^2 can be found. Also in this case, the weighted average calculated by (5.3.13) gives the maximum likelihood estimate of the white noise variance on the data series.

5.3.4 Conditional expectation value and covariance of $\boldsymbol{r}_j$

In the previous sections, the Maximum Likelihood re-estimation equations for each of the parameters $\theta = \{\mu_d, \sigma_d, \Phi, \boldsymbol{\varphi}, \sigma_\varepsilon^2\}$ were derived. However, to employ the re-estimation equations for Φ and $\boldsymbol{\varphi}$, we need to be able to evaluate the expectation value and covariance of $\boldsymbol{r}_j$ when conditioned on an observation segment $\boldsymbol{O}_j$, which is postulated to form an annual layer.

A simple way to derive these expectation values is to first consider the joint distribution of $\boldsymbol{O}_j$ and $\boldsymbol{r}_j$. As both $\boldsymbol{O}_j$ and $\boldsymbol{r}_j$ are Gaussian distributed with $\boldsymbol{O}_j \sim \mathcal{N}_{Md}(\mathrm{X}\boldsymbol{\varphi},\ \mathrm{X}\Phi\mathrm{X}^\intercal + \sigma_\varepsilon^2 \mathrm{W})$ and $\boldsymbol{r}_j \sim \mathcal{N}_K(\boldsymbol{0}, \Phi)$ (see section 4.2.3) also their joint distribution will be Gaussian. Most parameters of the joint distribution are directly given from the distributions of $\boldsymbol{O}_j$ and $\boldsymbol{r}_j$:

$$\begin{bmatrix}\boldsymbol{O}_j\\ \boldsymbol{r}_j\end{bmatrix}\sim\mathcal{N}_{KMd}\left(\begin{bmatrix}\mathrm{X}\boldsymbol{\varphi}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathrm{X}\Phi\mathrm{X}^{\intercal}+\sigma_\varepsilon^2\mathrm{W} & ?\\ ? & \Phi\end{bmatrix}\right)$$

The two yet unknown quantities in the joint covariance matrix can be calculated by:

$$\begin{aligned}\mathrm{cov}\left[\boldsymbol{O}_j,\boldsymbol{r}_j\right] &= \mathbb{E}\left[\left(\boldsymbol{O}_j-\mathbb{E}\left[\boldsymbol{O}_j\right]\right)\left(\boldsymbol{r}_j-\mathbb{E}\left[\boldsymbol{r}_j\right]\right)^{\intercal}\right]\\ &= \mathbb{E}\left[\left(\mathrm{X}\boldsymbol{r}_j+\boldsymbol{E}_j\right)\boldsymbol{r}_j^{\intercal}\right]\\ &= \mathrm{X}\mathbb{E}[\boldsymbol{r}_j\boldsymbol{r}_j^{\intercal}]\\ &= \mathrm{X}\Phi\end{aligned}$$

In the above, it was utilized that $\boldsymbol{O}_j$ is parameterized as $\boldsymbol{O}_j=\mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)+\boldsymbol{E}_j$, implying that $\boldsymbol{O}_j-\mathbb{E}\left[\boldsymbol{O}_j\right]=\mathrm{X}\boldsymbol{r}_j+\boldsymbol{E}_j$, where $\mathbb{E}\left[\boldsymbol{E}_j\right]=0$. Also, it was used that:

$$\mathbb{E}\left[\boldsymbol{r}_j\boldsymbol{r}_j^{\intercal}\right]=\ \mathbb{E}\left[\boldsymbol{r}_j\right]\mathbb{E}\left[\boldsymbol{r}_j\right]^{\intercal}+\mathrm{cov}\left[\boldsymbol{r}_j\right]=\mathrm{cov}\left[\boldsymbol{r}_j\right]=\Phi$$

Correspondingly (using that Φ is symmetric):

$$\mathrm{cov}\left[\boldsymbol{r}_j,\boldsymbol{O}_j\right]=\mathrm{cov}\left[\boldsymbol{O}_j,\boldsymbol{r}_j\right]^{\intercal}=(\mathrm{X}\Phi)^{\intercal}=\Phi^{\intercal}\mathrm{X}^{\intercal}=\Phi\mathrm{X}^{\intercal}$$

The joint distribution is therefore given as:

$$\begin{bmatrix}\boldsymbol{O}_j\\ \boldsymbol{r}_j\end{bmatrix}\sim\mathcal{N}_{KMd}\left(\begin{bmatrix}\mathrm{X}\boldsymbol{\varphi}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathrm{X}\Phi\mathrm{X}^{\intercal}+\sigma_\varepsilon^2\mathrm{W} & \mathrm{X}\Phi\\ \Phi\mathrm{X}^{\intercal} & \Phi\end{bmatrix}\right)\tag{5.3.14}$$

From the joint probability distribution, the conditional expectation value and covariance of $\boldsymbol{r}_j$ can be calculated as (a derivation is included in box 3):

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right] &= (\sigma_\varepsilon^2\Phi^{-1}+\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\mathrm{W}^{-1}\left(\boldsymbol{O}_j-\mathrm{X}\boldsymbol{\varphi}\right)\\ \mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right] &= \sigma_\varepsilon^2(\sigma_\varepsilon^2\Phi^{-1}+\ \mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X})^{-1}\end{aligned}$$

The conditional expectation value of $\boldsymbol{r}_j$, $\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$, is used for re-estimating $\boldsymbol{\varphi}$. The expectation value $\mathbb{E}\left[\boldsymbol{r}_j\boldsymbol{r}_j^{\intercal}|\ \boldsymbol{O}_j,\theta^{(k)}\right]$ employed in the update equation for Φ, can be calculated from the above as:

$$\mathbb{E}\left[\boldsymbol{r}_j\boldsymbol{r}_j^{\intercal}|\ \boldsymbol{O}_j,\theta^{(k)}\right]=\ \mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]^{\intercal}+\mathrm{cov}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$$

The conditioning on $\theta^{(k)}$ is included to clarify that the involved parameter values are to be taken from the current set of parameters.

Observe the difference between the conditioned and not-conditioned expectation values: The mean value of $\boldsymbol{r}_j$, $\mathbb{E}\left[\boldsymbol{r}_j\right]$, is zero, whereas the conditioned mean value, $\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$, generally is not. When conditioning on an observation segment, the expectation value is influenced by the observations and one obtains the value of the most likely random effect vector $\boldsymbol{r}_j$ for the layer in consideration.

Box 3: Conditional expectation and covariance of $\boldsymbol{r_j}$

Consider the general case of a joint probability distribution given as:

$$\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\right)$$

The mean and covariance of the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ can then be calculated by [*Bishop*, 2006]:

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] &= \boldsymbol{\mu}_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x)\\ \mathrm{cov}[\boldsymbol{y}|\boldsymbol{x}] &= \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\end{aligned}$$

Inserting the expressions for the mean and covariance of the joint distribution of $\boldsymbol{O}_j$ and $\boldsymbol{r}_j$ (5.3.14), we find for the conditional expectation value of $\boldsymbol{r}_j$:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j, \theta^{(k)}\right] &= \boldsymbol{0} + \Phi \mathrm{X}^{\intercal}(\mathrm{X}\Phi\mathrm{X}^{\intercal} + \sigma_\varepsilon^2\mathrm{W})^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\boldsymbol{\varphi}\right)\\ &= (\Phi^{-1} + \mathrm{X}^{\intercal}\sigma_\varepsilon^{-2}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\sigma_\varepsilon^{-2}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\boldsymbol{\varphi}\right)\\ &= (\sigma_\varepsilon^2\Phi^{-1} + \mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\intercal}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\boldsymbol{\varphi}\right)\end{aligned}$$

In the above, the Woodbury matrix identity has been used. For two positive definite matrices A and C, the general form of this matrix identity is:

$$(\mathrm{A} + \mathrm{BCD})^{-1} = \mathrm{A}^{-1} - \mathrm{A}^{-1}\mathrm{B}(\mathrm{DA}^{-1}\mathrm{B} + \mathrm{C}^{-1})^{-1}\mathrm{DA}^{-1}$$

Which can be re-arranged to [*Petersen and Pedersen*, 2008]:

$$(\mathrm{A} + \mathrm{BCD})^{-1}\mathrm{BC} = \mathrm{A}^{-1}\mathrm{B}(\mathrm{DA}^{-1}\mathrm{B} + \mathrm{C}^{-1})^{-1}$$

The latter version of the Woodbury identity was used in the above. The former version can be used for calculating the conditional covariance matrix:

$$\begin{aligned}\mathrm{cov}\left[\boldsymbol{r}_j|\boldsymbol{O}_j, \theta^{(k)}\right] &= \Phi - \Phi\mathrm{X}^{\intercal}(\mathrm{X}\Phi\mathrm{X}^{\intercal} + \sigma_\varepsilon^2\mathrm{W})^{-1}\mathrm{X}\Phi\\ &= (\Phi^{-1} + \mathrm{X}^{\intercal}\sigma_\varepsilon^{-2}\mathrm{W}^{-1}\mathrm{X})^{-1}\\ &= \sigma_\varepsilon^2(\sigma_\varepsilon^2\Phi^{-1} + \mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X})^{-1}\end{aligned}$$

5.3.5 Conditional expectation value of weighted residuals

The expectation values $\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j, \theta^{(k)}\right]$ and $\mathbb{E}\left[\boldsymbol{r}_j\boldsymbol{r}_j^{\intercal}|\,\boldsymbol{O}_j, \theta^{(k)}\right]$ used for the re-estimation of Φ and $\boldsymbol{\varphi}$ can be evaluated directly from a possible annual layer segment in the data series. This is not the case for the weighted squared residuals, whose expectation value is used in the update equation for σ_ε^2. With the residual vector given as $\boldsymbol{E}_j = \boldsymbol{O}_j - \mathrm{X}(\boldsymbol{\varphi} + \boldsymbol{r}_j)$, the magnitude of its components depends on the random effect of the specific layer ($\boldsymbol{r}_j$) as well as the mean layer trajectory parameter ($\boldsymbol{\varphi}$). Due to the dependency on the mean layer trajectory, the expectation value of squared residuals can only be calculated after the computation of an improved estimate of $\boldsymbol{\varphi}$.

The expectation value of the weighted sum of squared residuals, $\mathbb{E}\left[\boldsymbol{E}_j^{\intercal}\mathrm{W}^{-1}\boldsymbol{E}_j|\boldsymbol{O}_j, \theta^{(k)}\right]$, can be evaluated in terms of $\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j, \theta^{(k)}\right]$ and $\mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j, \theta^{(k)}\right]$ as follows:

$$\mathbb{E}\left[\boldsymbol{E}_j^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{E}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$$
$$\equiv \int_{\boldsymbol{r}_j\in\mathcal{R}} P\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\left(\boldsymbol{O}_j-\mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)\right)^{\mathrm{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j-\mathrm{X}\left(\boldsymbol{\varphi}+\boldsymbol{r}_j\right)\right)d\boldsymbol{r}_j$$
$$=\left(\boldsymbol{O}_j-\mathrm{X}\left(\boldsymbol{\varphi}+\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\right)^{\mathrm{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j-\mathrm{X}\left(\boldsymbol{\varphi}+\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\right)$$
$$+\mathrm{tr}\left(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$

where tr(·) signifies the trace. A derivation of this equality is found in appendix A4.2. Using the expectation value of $\boldsymbol{r}_j$ for the proposed layer to estimate its trajectory, the term $\boldsymbol{O}_j-\mathrm{X}\left(\boldsymbol{\varphi}+\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$ provides the resulting residuals. The first term in the equation above computes the weighted sum of the square of these. However, the uncertainties associated with estimating the appropriate random effect vector introduce uncertainties in the residuals hereby found. By adding the last term, these uncertainties are taken into account.

5.3.6 Finding the most likely annual layer parameters

Based on a single iteration of the Forward-Backward algorithm, a new and improved set of parameter values can be found from the equations (5.3.7), (5.3.8), (5.3.9), (5.3.12) and (5.3.13). The theory behind the EM-algorithm ensures these new parameter values to have a higher (or, at least, not a lower) likelihood than the original ones, and therefore to better describe the annual layers observed in the data series. By iteratively computing the layer segmentation probabilities with the Forward-Backward algorithm and updating the annual layer parameters, the algorithm is able to learn the appropriate parameter values, and it will converge towards a (local) maximum likelihood of these. The convergence criterion employed here will be defined as:

$$\left|\log L\left(\theta^{(k+1)}\middle|\boldsymbol{o}_{1:T}\right)-\log L\left(\theta^{(k)}\middle|\boldsymbol{o}_{1:T}\right)\right|<\epsilon,$$

with an appropriate choice of $\epsilon>0$.

The derived equations for re-estimating the different parameters used in the HMM layer detection model constitute a modified version of the set of equations used for waveform modeling by *Kim et al.* [2004] and *Kim and Smyth* [2006]. The re-estimation equations take into account both the uncertainty in layer positions as well as the uncertainty in parameter estimates based on the proposed segmentation of the data series into annual layers. The update equations can be interpreted in a sensible way, and one might have been able to guess some of them without turning to the math behind. Nevertheless, details in the resulting equations reveal the necessity of going through all these mathematical derivations in order to ensure their correctness. As an example, care must be taken to ensure that the right set of parameters ($\theta^{(k)}$ or the updated ones, $\theta^{(k+1)}$) are used in the calculations.

5.4 Maximum a Posteriori layer parameters

By means of the Forward-Backward algorithm, the most likely layering of the observation sequence can be found, and the algorithm may be trained by successive updates of the

model parameters used in the characterization of an annual layer. In this section, it will be described how such updates can be made to also take prior information into account.

The update equations derived in the previous section provide a Maximum Likelihood estimate of the model parameters. However, such estimates are only justified if the data series, on which the algorithm is trained, is sufficiently long to contain robust statistics on the parameter values. If limited to a relatively short observation sequence containing e.g. 30 years, the resulting parameter estimates are based on only 30 inferred annual layers. When using a relatively complex annual layer model, this is on the limit of providing accurate statistics for the model parameters, in particular for estimates of layer variance. This is most notably the case when considering data series with a high noise level, which generally require a larger body of data for correct assessment of the involved parameters. In such cases, the layering in the data itself may simply not be enough to constrain the model sufficiently for reliable Maximum Likelihood estimates to be made.

To stabilize the performance, the iterative improvement of model parameters employed in the Forward-Backward algorithm can be conducted in a Maximum a Posteriori (MAP) mode [*Gauvain and Lee*, 1994], which is less demanding on data quality and volume. In this case, prior knowledge on the individual parameter values is taken into account during their re-estimation. For annual layer detection in ice cores, such prior information may consist of knowledge derived from previous data on how the annual layers generally appear in the observation sequence, as well as an estimate of the annual layer thickness distribution in the preceding depth interval. Incorporating such information in the training process increases the stability of the algorithm, and estimations should be more robust for short observations sequences.

Prior knowledge on parameters is incorporated into the layer detection algorithm in form of prior probability distributions for the individual parameter values. These prior probability distributions are described by hyper-parameters (which are not to be confused with the model parameters themselves). Such prior information going into the model may be as complex as desired, and entirely depends on the relevant assumptions for the problem at hand. Here, we have focused on the simplest one. To facilitate the ensuing analysis, the prior for the duration parameter μ_d is chosen as a normal distribution described by the two hyper-parameters m_μ and v_μ, i.e. $\mu_d \sim \mathcal{N}(m_\mu, v_\mu)$. Similarly, the prior for the parameter describing the mean annual layer shape, $\boldsymbol{\varphi}$, is taken as a multivariate normal distribution with mean $\boldsymbol{u}_\varphi$ and covariance matrix U_φ, i.e. $\boldsymbol{\varphi} \sim \mathcal{N}_K(\boldsymbol{u}_\varphi, \mathrm{U}_\varphi)$. In order not to increase the complexity further, the remaining parameters are considered fixed and known. The reasoning behind this choice of fixed versus adaptable parameters will be discussed in section 6.2.1.

The collection of hyper-parameters will be denoted by Θ, and comprises the following: $\Theta = \{m_\mu, v_\mu, \sigma_d^{(0)}, \boldsymbol{u}_\varphi, \mathrm{U}_\varphi, \Phi^{(0)}, \sigma_\varepsilon^{(0)}\}$. Parameters, which are assumed known, can be regarded as having deterministic priors, and the parameter and hyper-parameter is equivalent: $\sigma_d = \sigma_d^{(0)}$, $\Phi = \Phi^{(0)}$, and $\sigma_\varepsilon = \sigma_\varepsilon^{(0)}$. The set of hyper-parameters contains all prior information on model parameters employed in the Forward-Backward algorithm for annual layer detection. Accordingly, it includes all information required for iterative MAP-updates of these parameters.

Given the assumption of known model parameters σ_d, Φ, and σ_ε, the resulting update equations will not represent a full generalization to the case of Maximum a Posteriori parameter re-estimates. Keeping the above parameters fixed, they do not need to be re-estimated, hence necessitating the use of fewer update equations than derived for the Maximum Likelihood in the previous section. In other respects, however, the Maximum a Posteriori methodology does represent a more complex approach: A single adaptable parameter in the Maximum Likelihood update equations is now being described by two new hyper-parameters. Furthermore, these hyper-parameters must be adjusted prior to the layer detection procedure being carried out. Fortunately, they can usually be estimated based on previously processed data, hence limiting the number of subjective tuning parameters.

5.4.1 Layer thickness parameters

The annual layer thicknesses are taken to be distributed according to a lognormal distribution with location parameter μ_d and scale parameter σ_d. As mentioned above, the scale parameter of the distribution will in the following be assumed known. The prior for the location parameter is chosen as a normal distribution:

$$\mu_d \sim \mathcal{N}(m_\mu, v_\mu), \;\; \sigma_d \; = \sigma_d^{(0)}$$

That is:

$$P(\mu_d) = \frac{1}{\sqrt{2\pi v_\mu}} \exp\left(-\frac{(\mu_d - m_\mu)^2}{2v_\mu} \right)$$

To obtain a Maximum a Posteriori estimate of μ_d, this prior must be used in the M-step of the EM-algorithm, which is now a maximization of the auxiliary function $R(\theta|\theta^{(k)})$ (5.2.2). The maximization of this function is done by differentiating with respect to the parameter μ_d, and setting the derivative equal to zero.

The derivative of the function $R(\theta|\theta^{(k)}) = Q(\theta|\theta^{(k)}) + \log P(\theta)$ can be rewritten in terms of the derivative of the Q-function:

$$\frac{\partial R(\theta|\theta^{(k)})}{\partial \mu_d} = \frac{\partial Q(\theta|\theta^{(k)})}{\partial \mu_d} + \frac{\partial \log P(\theta)}{\partial \mu_d}$$

It was previously shown that (5.3.6):

$$\frac{\partial Q(\theta|\theta^{(k)})}{\partial \mu_d} = \frac{1}{\sigma_d^2} \sum_{t,j,d} \bar{\eta}_t(j,d)(\log d - \mu_d)$$

Inserting this, along with the derivative of the assumed prior for μ_d, into the expression for the derivative of the R-function, yields:

$$\frac{\partial \log P(\theta)}{\partial \mu_d} = \frac{\partial \log P(\mu_d)}{\partial \mu_d} = \frac{\partial}{\partial \mu_d}\left(-\frac{1}{2}\log 2\pi v_\mu - \frac{(\mu_d - m_\mu)^2}{2v_\mu} \right) = -\frac{\mu_d - m_\mu}{v_\mu}$$

$$\frac{\partial R(\theta|\theta^{(k)})}{\partial \mu_d} = \frac{1}{\sigma_d^2}\sum_{t,j,d} \bar{\eta}_t(j,d)(\log d - \mu_d) - \frac{\mu_d - m_\mu}{v_\mu}$$

Equating this expression with zero, leads to the following Maximum a Posteriori update equation for the parameter μ_d:

$$\frac{1}{\sigma_d^2}\sum_{t,j,d} \bar{\eta}_t(j,d)(\log d - \mu_d) - \frac{\hat{\mu}_d - m_\mu}{v_\mu} = 0 \;\Leftrightarrow$$

$$\boxed{\hat{\mu}_d = \frac{v_\mu \sum_{t,j,d} \bar{\eta}_t(j,d)\log d + \sigma_d^2 m_\mu}{v_\mu \sum_{t,j,d} \bar{\eta}_t(j,d) + \sigma_d^2}} \qquad (5.4.1)$$

Note that this re-evaluation of μ_d not only depends on the hyper-parameters m_μ and v_μ describing the prior distribution of μ_d. It is also influenced by the scale parameter of the layer thickness distribution (σ_d), which here has been assumed known. Consider the case where the value of σ_d is known to be very large (e.g. $\sigma_d \to \infty$). In this situation, we cannot have much faith in the estimate of μ_d derived from a weighted sample average of the data. As a consequence, the a posteriori most probable value of μ_d is almost completely determined by the prior, and $\hat{\mu}_d \approx m_\mu$.

The case of no prior knowledge (equivalent to the maximum likelihood case) can be obtained as a special case of the above. A non-informative prior corresponds to letting the variance of the prior probability distribution for μ_d approach infinity: $v_\mu \to \infty$ (and hence $\sigma_d^2/v_\mu \to 0$). In this case, the update equation for μ_d becomes similar to the one derived in section 5.3.1 for the maximum likelihood case. Conversely, in case of a very constraining prior ($v_\mu \to 0$), the weighted sample average obtained from data has almost no influence, and $\hat{\mu}_d \approx m_\mu$.

In this way, (5.4.1) works by balancing the two terms: The weighted sample average of the location parameter of the layer thickness distribution as found from data, and the prior probability distribution for μ_d. The resulting value of μ_d depends on the constraints imposed by our prior knowledge, and the degree to which data is believed to contain information on the layer thickness distribution.

5.4.2 Annual layer signal parameters

For the Maximum a Posteriori update equations developed here, most of the layer trajectory parameters are assumed to be known beforehand. Only the mean layer signal parameter, $\boldsymbol{\varphi}$, will be re-estimated based on data, whereas both Φ and σ_ε are considered fixed. A further discussion on this matter can be found in section 6.2.1.

The mean layer trajectory vector, $\boldsymbol{\varphi}$, generally contains more than just a single value. To avoid excessive complexity, the prior distribution for this trajectory parameter is here taken to be a multivariate normal distribution with mean vector $\boldsymbol{u}_\varphi$ and covariance matrix U_φ. Hence, the complete prior for the annual layer signal parameters is as follows:

$$\boldsymbol{\varphi} \sim \mathcal{N}_K(\boldsymbol{u}_\varphi, \mathrm{U}_\varphi), \quad \Phi = \Phi^{(0)}, \quad \sigma_\varepsilon = \sigma_\varepsilon^{(0)}$$

K is the number of parameters used for modeling an annual layer signal.

The difference between the two covariance matrices U_φ and Φ deserves a short comment. Both matrices describe how the layer signal is allowed to change. However, the fixed covariance matrix Φ states the allowed variability of annual layers around their mean trajectory, whereas the covariance matrix U_φ describes the uncertainty on this mean signal. Thus, the information they contain is not redundant.

As before, the update equations for $\boldsymbol{\varphi}$ is found by maximizing the R-function (5.2.2) with respect to $\boldsymbol{\varphi}$. This task can be divided up into finding the derivative of the Q-function and of the prior log-probabilities. The derivative of the Q-function was derived in (5.3.11):

$$\frac{\partial Q\left(\theta\middle|\theta^{(k)}\right)}{\partial \boldsymbol{\varphi}} = \frac{1}{\sigma_\varepsilon^2}\sum_{t,j,d} \bar{\eta}_t(j,d)\,\mathrm{X}^\intercal \mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi} + \mathbb{E}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\right)$$

The prior probability of a given value of $\boldsymbol{\varphi}$ can be written as:

$$P(\boldsymbol{\varphi}) = (2\pi)^{-K/2}\left|\mathrm{U}_\varphi\right|^{-1/2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)^\intercal \mathrm{U}_\varphi^{-1}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)\right)$$

And accordingly, the derivative of the prior log-probabilities is given by:

$$\begin{aligned}\frac{\partial \log P(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} &= \frac{\partial}{\partial \boldsymbol{\varphi}}\left(-\frac{K}{2}\log 2\pi - \frac{1}{2}\log\left|\mathrm{U}_\varphi\right| - \frac{1}{2}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)^\intercal \mathrm{U}_\varphi^{-1}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)\right)\\ &= \frac{\partial}{\partial \boldsymbol{\varphi}}\left(-\frac{1}{2}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)^\intercal \mathrm{U}_\varphi^{-1}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)\right)\\ &= -\mathrm{U}_\varphi^{-1}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)\end{aligned}$$

The last equality can be derived by completing the squares, and differentiating each term separately, while using that the covariance matrix U_φ is symmetric. The derivation is almost analogous to the one included in appendix A4.1.

Inserting these two into the expression for the derivative of the R-function, we arrive at the following:

$$\begin{aligned}\frac{\partial R(\theta|\theta^{(k)})}{\partial \boldsymbol{\varphi}} &= \frac{\partial Q\left(\theta\middle|\theta^{(k)}\right)}{\partial \boldsymbol{\varphi}} + \frac{\partial \log P(\theta)}{\partial \boldsymbol{\varphi}}\\ &= \frac{1}{\sigma_\varepsilon^2}\sum_{t,j,d} \bar{\eta}_t(j,d)\,\mathrm{X}^\intercal \mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi} + \mathbb{E}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\right) - \mathrm{U}_\varphi^{-1}\left(\boldsymbol{\varphi}-\boldsymbol{u}_\varphi\right)\end{aligned}$$

Equating this with zero, the following MAP-update equation for the mean layer signal parameter vector is obtained:

$$\boxed{\widehat{\boldsymbol{\varphi}} = \left[\sum_{t,j,d} \bar{\eta}_t(j,d)\,\mathrm{X}^\intercal \mathrm{W}^{-1}\mathrm{X} + \sigma_\varepsilon^2 U_\varphi^{-1}\right]^{-1}\left(\sum_{t,j,d} \bar{\eta}_t(j,d)\,\mathrm{X}^\intercal \mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\mathbb{E}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right) + \sigma_\varepsilon^2 U_\varphi^{-1}\boldsymbol{u}_\varphi\right)}$$

(5.4.2)

As was the case for the re-estimated value of μ_d (5.4.1), also this Maximum a Posteriori re-evaluation of the mean layer signal parameter represents a weighting between its sam-

ple average and its prior distribution. The weighting depends on the relative amount of information contained in the two. As the remaining parameters in our case are kept fixed, only these two update equations are required.

By utilizing the derived MAP update equations for μ_d and $\boldsymbol{\varphi}$, while keeping fixed all remaining parameters used in the Forward-Backward algorithm, an improved assessment of the value of these two parameters can be made. The re-evaluation takes into account our prior knowledge on the value of these variables. By iteratively estimating a new and better set of model parameters, and subsequently running the Forward-Backward algorithm based on these, a joint set of parameters with maximal posterior probability will eventually be found. However, the resulting Maximum a Posteriori parameter estimates are of course dependent on the postulated fixed values of σ_d, Φ, and σ_ε, as well as on the assumed prior for the adjustable model parameters μ_d and $\boldsymbol{\varphi}$. And, as it is always the case with results based on the EM-algorithm, the located maximum may just be a local maximum.

5.4.3 Posterior probability of the joint set of parameters

The successive iterations of the EM-algorithm are terminated when the algorithm has managed to converge to a local maximum of the posterior probability function for the joint set of model parameters θ. When running the EM-algorithm in Maximum a Posteriori mode, the convergence criterion may be defined as:

$$\left|\log P\left(\theta^{(k+1)}\middle|\boldsymbol{o}_{1:T}\right) - \log P\left(\theta^{(k)}\middle|\boldsymbol{o}_{1:T}\right)\right| < \epsilon$$

for an appropriate choice of $\epsilon > 0$.

To determine whether or not this convergence criterion has been reached, the posterior probability of the model parameters must be assessed. This probability can be evaluated based on the obtained likelihood of the model parameters and their prior distributions. The likelihood of the model parameters, $L\left(\theta^{(k)}\middle|\boldsymbol{o}_{1:T}\right) = P(\boldsymbol{o}_{1:T}|\theta^{(k)})$, is derived directly during the computations of the Forward-Backward algorithm. The prior probabilities of the individual model parameters are known beforehand. The posterior log-probabilities of the current set of model parameters can then be calculated as follows:

$$\begin{aligned}\log P\left(\theta^{(k)}\middle|\boldsymbol{o}_{1:T}\right) &\propto \log P\left(\boldsymbol{o}_{1:T}\middle|\theta^{(k)}\right)P\left(\theta^{(k)}\right)\\ &= \log P\left(\boldsymbol{o}_{1:T}\middle|\theta^{(k)}\right) + \log P\left(\theta^{(k)}\right)\\ &= \log P\left(\boldsymbol{o}_{1:T}\middle|\theta^{(k)}\right) + \log P\left(\mu_d^{(k)}\right) + \log P\left(\boldsymbol{\varphi}^{(k)}\right)\end{aligned}$$

The proportionality constant, $P(\boldsymbol{o}_{1:T})$, remains the same for all iterations, and can therefore be ignored. The values of $\log P\left(\mu_d^{(k)}\right)$ and $\log P\left(\boldsymbol{\varphi}^{(k)}\right)$ are assessed based on their prior probability distributions.

5.5 Improvement of parameters

The most likely annual layering in an observation sequence as determined by the Forward-Backward algorithm (or the Viterbi algorithm) depends on the annual layer model and

model parameters used as input to the algorithm. However, the dependency of the result on the employed model parameters can be either partly or completely alleviated. This is done by training the algorithm on the observed data, thereby allowing it to use the joint set of layer model parameters which fit the observations 'best'. Such training of the algorithm can be achieved as the likelihood of the employed layer model parameters based on the data is calculated directly during the Forward-Backward algorithm procedure.

The EM-algorithm presents a relatively simple and fast way of obtaining 'best' estimates of the model parameters used for describing an annual layer in the observation sequence. In addition, the algorithm can be adapted to allow such estimates to take into account prior information on the layer model parameters, as it can be run in both a Maximum Likelihood (without prior) and a Maximum a Posteriori (with prior) mode.

If running the EM-algorithm in Maximum Likelihood mode, the resulting most likely annual layering of the observation sequence is (almost) completely independent on the input parameters used. The slight dependency which remains is due to the deterministic behavior of the EM-algorithm which may cause it to get caught up in a local maximum. To avoid this, the algorithm may be run multiple times with different starting points, hereby increasing the chances of finding the set of layer model parameters having globally maximum likelihood.

However, for annual layer detection in ice core data, a more constrained version of the EM-algorithm may be required. We have no perfect model for the expression of an annual layer in the data, and this may cause an unconstrained version of the EM-algorithm to go completely astray. Furthermore, even with a perfect annual layer model, the use of relatively short data sequences containing perhaps only 30 years may not be sufficient to produce reliable Maximum Likelihood layer parameter estimates. In this case, prior information on the layer model parameters may be taken into account, and stabilize the methodology. In many ways, the Maximum a Posteriori approach is very beneficial. Yet, Maximum a Posteriori estimates do per definition depend on the employed prior. In practice, however, a reasonable estimate for such priors can often be made in an objective way based on inferred or observed layering in previous data intervals.

6. Layer detection in sequential batches of data

It is not feasible, nor desirable, to run the Forward-Backward algorithm and/or the Viterbi algorithm on perhaps several hundred meters of ice core data at once. While also increasing the computational complexity drastically, doing so would require a homogeneous data series, in which the annual layer thickness distribution as well as the layer signal is more or less constant. Neither of these conditions are satisfied: Annual layer thicknesses are changing down the ice core as the combined result of climate-induced variations in past accumulation rates and a general thinning of layers with depth due to ice flow. In different climate regimes, also the influx of impurities to the inner part of the ice sheet may differ – in quantity as well as seasonality – hereby altering the general annual layer signal in the ice core data.

A much better strategy is to divide the total data series into smaller batches, apply the layer detection algorithm to one of these at a time, and subsequently stitch them together. To retrieve the annual layering down the ice core in the best possible way, one must choose an appropriate length of such data batches. This length must be chosen such as to balance between the need of the observation sequences to be sufficiently long to fully exploit the HMM's optimal estimation of layer boundaries, while being short enough that the assumption of a fixed layer thickness distribution and layer signal is reasonable. Also, the shorter the length of these batches, the more efficiently does the algorithm run: The layer detection algorithm is linear (as opposed to exponential) in T and J (see section 3.4). However, increasing the length of an observation sequence (T), also requires an augmented maximum number of allowed annual layers in the sequence (J). The combined effect is a significant increase in computational burden.

Here, the algorithm has been chosen to run on batches of data covering approximately 30-50 years each, with the length of each batch being individually determined based on a first guess of the mean annual layer thickness. For the visual stratigraphy data within the selected depth interval, this amounts to 350-700 observations per batch. The choice of an approximately fixed number of annual layers within each batch, instead of e.g. a fixed

batch length, was made to ensure the control over the number of layers on which the re-estimated parameter values are based.

During fast climatic shifts, however, the mean annual layer thickness may change within a very short time period indeed. At the onset of the Holocene, two abrupt warming events occurred, interrupted by the Younger Dryas cold period. During the warming events, the annual layer thicknesses increased by 40% over respectively 3 and 40 years. The change in layer thicknesses happened slightly slower during the intermediate cooling event (i.e. the onset of Younger Dryas), during which the mean annual layer thickness decreased with 33% over 152 years [*Steffensen et al.*, 2008].

Nevertheless, the length of the data series cannot be much further reduced, and for most time periods, it is a reasonable assumption that the accumulation rates will not change significantly during a 30-50 year epoch. Meanwhile, even if 30-50 layers may be too few to provide viable Maximum Likelihood estimates of the annual layer parameters, it should still be possible to obtain robust parameter estimates if prior information on the parameter values is included.

6.1 Combining successive data batches

By dividing the data series into batches, and running the layer detection algorithm on each of these individually, some of the information contained in the full data series is lost. Close to both edges of each batch, the lacking knowledge on the surrounding data outside the batch will in general cause the annual layer boundaries here to be placed less accurately. To some extent, however, such knowledge can be recovered by choosing data batches in consecutive order, and incorporating some of the information inferred from one batch of data into the next.

To initialize the Forward-Backward/Viterbi algorithm, information on starting position of the first layer in the current batch is utilized. Such information may just be the common-sense logic that the very first layer started somewhere before the first observation in the data batch. In that case, the general initialization condition (3.4.6) can be applied. Yet, the performance of the layer detection algorithm will of course improve when adding more detailed information on the probability distribution corresponding to the position of the preceding layer boundary (3.4.7). Such knowledge can be obtained based on the most likely annual layering, as inferred by the Forward-Backward algorithm, in the last part of the previous batch of data.

However, the layering in the very last part of each batch of data is generally less accurately determined than the rest. An additional initialization condition for the layer detection algorithm is the information on ending position of the very last layer in the observation sequence. With no such information available, the general condition of the last layer ending somewhere after the last observation in the sequence, is applied (3.4.9). As a consequence of this unconstrained initialization condition, the quality of the inferred layering generally degrades towards the end of each observation sequence. To lessen the importance of this issue, the last part of each batch is discarded after having been used for inferring a best estimate of the annual layering in the complete batch.

In this way, the layer detection algorithm can be run on successive, slightly overlapping batches of data. Only the first part of the inferred layering of each batch is accepted. The final part is rejected. The annual layering here can be better determined when also including the information contained within subsequent data. The initial condition for the next batch is then determined based on the deduced most likely layering in the last portion of the accepted part of the observation sequence. In this manner, knowledge based on previous data is continually being incorporated into the analysis of the following data batch, thereby minimizing the deterioration in quality of the reconstructed layering caused by batch edges.

6.1.1 Shortening the observation sequence

Wishing to run the annual layer detection algorithm in the overlapping fashion described above, it must, after the analysis of each batch of data, be decided how much of the observation sequence should be discarded. This should be done for that part of the observation sequence for which a much better estimate of the annual layering would be made if also taking data subsequent to the present batch into consideration.

The required length of the overlapping section depends on the data in question. Without the surrounding data, the annual layer boundaries in the last part of the batch are generally determined with less certainty. If, in spite of this, one of the last layer boundaries in the observation sequence is very well-defined, it is sufficient to discard only the very last part of the batch. Conversely, if the layering in the last part of the observation sequence is ambiguous, it may be necessary to discard a relatively large fraction of the current batch of data.

The Forward-Backward algorithm provides an estimate of the certainty with which the individual layer boundaries have been determined, both in terms of their positional accuracy as well as their associated certainty of being a layer boundary. We wish to make use of this knowledge to find a very well-defined layer boundary in the last part of the observation sequence. As such, it does not bear much importance whether or not the layer boundary is accurately positioned or not. What matters is that it is very certain to be a layer boundary. However, the two are not unrelated: At a location where the probability of having a layer boundary is high, this is most likely a very certain layer boundary.

Hence, an indication of a certain layer boundary is a high probability of ending any layer ℓ_j at a given t. To obtain the value of t with the highest such probability, the following quantity is maximized (figure 6.1.1):

$$t_b = \underset{t \in \mathcal{T}_b}{\operatorname{argmax}} \sum_{\ell_j \in \mathcal{L}} P\left(S_{t]} = \ell_j, \boldsymbol{o}_{1:T}\right) = \underset{t \in \mathcal{T}_b}{\operatorname{argmax}} \sum_{j=1}^{J} \sum_{d=1}^{D} \eta_t(j,d)$$

$\mathcal{T}_b$ is an appropriate interval within the last part of the data sequence. This interval should be large enough to contain at least one well-defined layer boundary. On the other hand, in order to make sure that the layer, whose boundary we are considering, has ended before the start of the next batch of data, it should not contain the very last part of the observation sequence either. Here, $\mathcal{T}_b$ has been chosen as the interval between approximately 1-5 annual layer thicknesses before the end of the observation sequence.

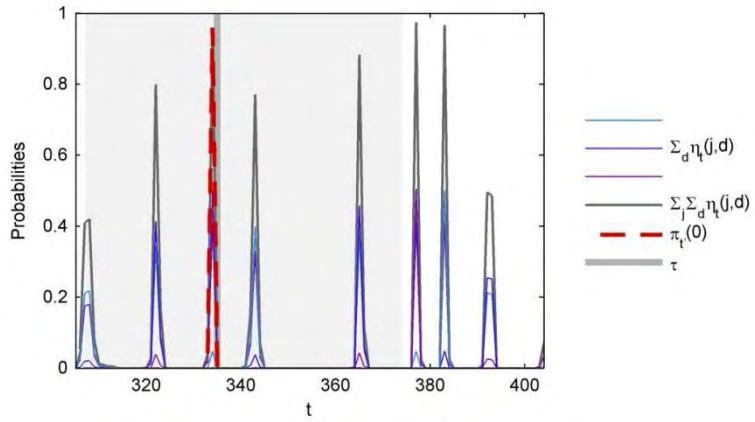


Figure 6.1.1: *Combining consecutive batches. Example from a depth of 2233 m. The shaded area represents the interval $\mathcal{T}_b$, within which the best cut-off location (τ) must be selected. It is chosen just after the spot with the highest probability of containing a layer boundary. The resulting probability distribution, which is to be used as initial condition for the subsequent batch, is shown in red.*

Above, the most likely position of a very certain layer boundary contained within $\mathcal{T}_b$ was computed. The cut-off point of the current data batch, τ, is then chosen as the t encountered hereafter with the least (and expected: zero) probability of containing an annual layer boundary:

$$\tau = \underset{t_b<t\leq t_b+D}{\operatorname{argmin}} \sum_{j=1}^{J}\sum_{d=1}^{D} \eta_t(j,d)$$

In case of several occurrences of the minimum value, the first of these is chosen. At this value of t, we can be very confident that a new layer has just started, and that this new layer had a well-defined first layer boundary. Hence, an accurate initial condition for the next batch can be provided.

6.1.2 Initial condition for next batch

Having established τ to be a good choice for the cut-off position of the current batch of data, the initial conditions for a subsequent batch having this starting point must be determined. For initializing the forward pass (see (3.4.6)), the probabilities corresponding to the termination of the layer prior to the one in τ must be evaluated. Such probabilities can be computed as:

$$\pi_{t'}(0) = \sum_{\ell_j\in\mathcal{L}} P\big(S_{[t+1:\tau}=\ell_j\big) = \sum_{\ell_j\in\mathcal{L}}\sum_{\substack{d\geq\tau-t\\ d\in\mathcal{D}}} P\big(S_{[t+1:t+d]}=\ell_j\big)$$

Disregarding the conditioning on observations contained in the probability measure $\bar{\eta}_t(j,d)$, this can be approximated by:

$$\pi_{t'}(0) \approx \sum_{j=1}^{J}\sum_{d=\tau-t}^{D} \bar{\eta}_{t+d}(j,d)$$

The above initial condition, which describe the probabilities of the starting position of what is to become the new layer ℓ_0 of the subsequent batch, has here been indexed with t'

corresponding to this new batch: $t' = t - \tau + 1$. It provides the complete set of initial conditions required for accurate annual layer detection in the subsequent batch, which starts in τ.

When using this procedure for combining consecutive data batches, the resulting annual layer count was found to be almost identical to that resulting from running the algorithm with all parameters fixed on a complete section at once.

6.1.3 Resulting number of annual layers

The annual layers in the data series are always counted from the beginning of each batch, starting with ℓ_0. This is done in order to always keep the number of possible states in the system to a minimum. As the computational burden of the annual layer detection algorithm scales linearly with the number of states (section 3.4), an up-scaling of the complexity of the problem with increasing batch number is hereby avoided. However, it is the combined layer count based on the layering in all batches which is desired. Such a chronology down the ice core can be obtained by convolving the resulting layer probability distributions from each batch. In this manner, all information in the resulting annual layer counts and corresponding uncertainties can be retained with minimal amount of effort.

Denote as $\gamma_t^{total}(j)$ the merged probability of being in layer ℓ_j at t, given the entire collection of observations up to t. The change in indexing to t and j is made to clarify that these variables now are measured from the depth at which the layer counting algorithm was initiated. In contrast, the variables t and j are measured from the beginning of each batch. The probability measure $\gamma_t^{total}(j)$ contains the complete information on the resulting layer counted chronology down the ice core to t.

Consider a data batch k, which starts at t_k. Assume the merged probability distribution of counted annual layers corresponding to the very first observation in this batch to be known. This probability distribution is $\gamma_{t_k}^{total}(j)$. By use of the Forward-Backward algorithm, the probability distribution of the number of annual layers in this current data batch is given as $\bar{\gamma}_t(j)$, $t \leq \tau$. The conditioning of these probabilities on the observations in the batch will here be neglected. The merged probability distribution of annual layers throughout data batch k can then be calculated as the convolution of these two probability distributions:

$$\gamma_{t_k+t-1}^{total}(j) = \sum_{j=1}^{J} \gamma_{t_k}^{total}(j - j - 1) \cdot \bar{\gamma}_t(j)$$

And by the cut-off position of data batch k (i.e. for $t = \tau$), the merged probability distribution, as required for computing the merged probability distribution of the subsequent batch, is given by:

$$\gamma_{t_{k+1}}^{total}(j) = \gamma_{t_k+\tau-1}^{total}(j)$$

The probability distribution $\gamma_t^{total}(j)$ can be summarized using descriptive statistics such as the mean, median, quantiles etc. of the distribution. As for the case of a single batch

(section 3.5), these quantities can then be used for describing the corresponding layer counted ice core chronology and its associated uncertainties.

6.2 Changing parameter values down the core

One of the main reasons to split up the data series into batches, and perform the annual layer detection on each batch separately, is to allow the parameter values describing an annual layer to vary down the ice core. Provided that the parameters are varying sufficiently slowly, their values at any particular depth are relatively well-constrained from previous data. A major increase in performance may result from taking such knowledge into account. Approximate posterior probability distributions of model parameters derived from previous data can be applied as prior probabilities for the current batch, and in this way be used for constraining the parameter values here. By incorporating such knowledge, the algorithm is allowed to continuously adjust itself to changes in how an annual layer is expressed in the data series, and it does so in a flexible, yet controlled manner.

Each batch is chosen to contain around 30-50 annual layers. Within each of these batches, the layer thickness distribution and the annual layer signal are assumed constant. However, as the layer detection algorithm allows for a range of variability around the mean of both of these (the allowed amount being specified as a model parameter itself), the algorithm may still be able to pick up a slow evolution of the parameter values even within a batch. The requirement for this to happen is that the change in mean value is small compared to the allowed variability from one year to the next.

When using results based on previous data as prior for current data, the validity of imposing such prior is contingent on the layer model parameters to be slowly varying over time. The spread of their respective prior distributions determines the abruptness of changes allowed. The smaller the spread, the more constrained is the model, and the slower must the evolution take place.

The updating of priors for the parameter values from one batch to the next can be made in a variety of ways, one more sophisticated than the other (see e.g. [*Jen-Tzung Chien*, 2002; *J.-T. Chien and Huang*, 2003; *Gauvain and Lee*, 1994; *Huo and Lee*, 1997]). The most sophisticated methods may e.g. incorporate the dependency of the priors on each other, while the less pretentious ones settle for very simple and independent prior probability distributions. Also, some of the parameters may be considered fixed and known beforehand.

For the layer detection algorithm, a first step before including such prior information is to consider which model parameters can be kept fixed, and which ones must be allowed to adapt to the changing climate. By tying some of the parameters to a fixed value, and describe the remaining ones with very simple probability distributions, the tractability of the problem is greatly enhanced, and the changing characteristics of an annual layer in the data series can be traced down the ice core.

6.2.1 Adaptable and tied parameters

In section 5.4, it was explained how the EM-algorithm can be run in Maximum a Posteriori mode, thereby allowing prior knowledge on the parameter values to be taken into account. The update equations (5.4.1) and (5.4.2) were derived for a simplified case, in which three of the annual layer parameters (σ_d, Φ, and σ_ε^2) are considered known beforehand. The remaining two parameters (μ_d and $\boldsymbol{\varphi}$) are allowed to adapt themselves to the data, and their prior probability distributions are considered to belong to the family of normal distributions.

For the current application concerning annual layer detection in ice core data, it is fairly reasonable to assume the values of σ_d, Φ, and σ_ε^2 to be constant. Nor is the choice of prior probability distributions for μ_d and $\boldsymbol{\varphi}$ as normal distributions unreasonable. The arguments behind making such simplifying choices are described below.

Annual layer thickness parameters

The assumption of a known scale parameter of the layer thickness distribution (σ_d) can be justified based on previous studies, which indicate this parameter to be only marginally dependent on climate [*Andersen et al.*, 2006b]. Nor does the gradual ice-flow induced thinning of annual layers with depth contribute to changes in the general shape of the layer thickness probability distribution (section 4.1). Hence, it is not unreasonable to assume the scale parameter to maintain a rather constant value with depth.

However, when going into details, it is not clear how valid this assumption is. In figure 6.2.1C&D, the evolution with depth of the two layer thickness distribution parameters is shown. To make it resemble the outcome from successive batches of data, the statistics of these parameters is based on 50 years each. It must be emphasized that the derived evolution of σ_d (figure 6.2.1D) very much depends on the exact placement of the GICC05 layer boundaries. Yet, the GICC05 chronology was not developed with the purpose of obtaining the best layer boundaries, but rather to obtain the best timescale. Most of the obtained variation in σ_d may therefore be artifacts due to e.g. the person in charge of the layer counting, the seasonal variability of peak events in the employed data series, as well as the general degree of difficulty in doing such counting.

From figure 6.2.1D, it is seen that the estimated value of σ_d does show some variation with depth. Sections of unusually large variations of individual layer thicknesses, i.e. large values of σ_d (marked as gray), tend to be associated with a decrease in mean layer thicknesses, a sign of climatic cooling events. However, there is no straightforward relation between the two. Many cooling events do not have a counterpart in σ_d, and not all sections of high σ_d values appear to be connected with variations in the $\delta^{18}O$ profile. Again, it may just be due to artifacts. Furthermore, σ_d seems to decrease with depth. This can possibly be explained by a changing counting strategy due to the general smoothing of the data series with depth. In any case, however, the variations in σ_d (figure 6.2.1 D) are much less pronounced than the changes in e.g. μ_d (figure 6.2.1B), and in the following, σ_d will be assumed constant with depth.

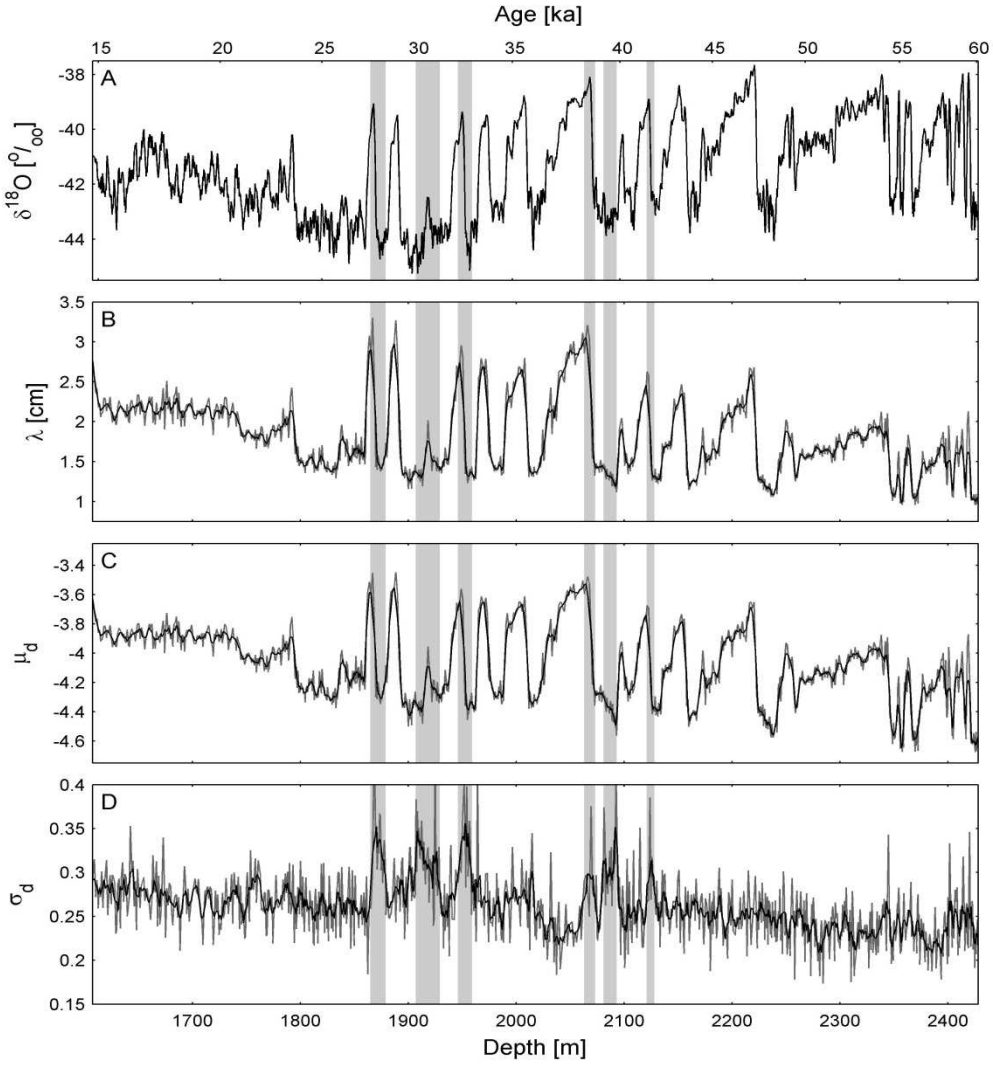


Figure 6.2.1: *Estimated evolution of the layer thickness distribution parameters μ_d and σ_d (C&D) for the lower part of the NGRIP ice core, and their relation to the observed $\delta^{18}O$ variations (A). Also the mean annual layer thickness, λ, as derived from μ_d and σ_d is shown (B). Layer thicknesses are based on the GICC05 chronology, and uncertain layer boundaries have not been included in the statistics. Each estimated value of μ_d and σ_d is based on 50 layers. Segments with unusually high values of σ_d are marked in gray.*

The mean annual layer thickness does change with depth, and sometimes quite abruptly. These abrupt changes are caused by changing climate conditions and the accompanying variations in accumulation rate, and can be seen as shifts in the location parameter of the annual layer thickness distribution (figure 6.2.1C). Furthermore, gradual changes in the location parameter result from thinning of the annual layers with depth due to ice flow. Hence, the location parameter of the annual layer thickness distribution cannot be assumed constant.

Annual layer thicknesses usually change gradually. In the upper part of the ice core, firn compaction and high strain rates cause the annual layer thicknesses to rapidly decrease

with depth. At these depths, a symmetrical distribution, such as e.g. the normal distribution, would not be a good description for the changes in layer thicknesses (as described by μ_d) from one batch of data to the next.

At larger depths, lower strain rates implies less impact on annual layer thicknesses from ice flow induced thinning, and higher impact from changes in climate. This is in particular the case for the deeper part of the NGRIP ice core. The occurrence of bottom melt at this location is reflected in very low thinning rates of annual layers with depth [*D Dahl-Jensen et al.*, 2002]. As a result, for the depth interval in consideration, the chance of layer thicknesses being smaller/larger in a subsequent batch is fairly equal, and it turns out that a normal distribution here is able to describe these changes quite well (figure 6.2.2B).

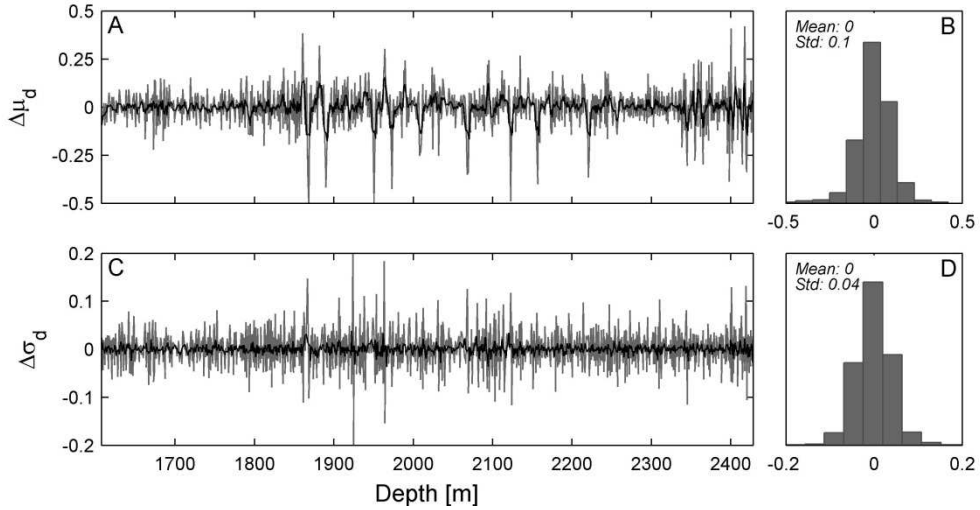


Figure 6.2.2: *The changes in layer thickness distribution parameters, σ_d and μ_d, from one "batch" of 50 layers to the next (A,C), and the resulting frequency distributions (B,D). Annual layer thickness parameters are derived from GICC05 layer positions, uncertain years not included.*

However, at extremely fast transition periods, where the annual layer thicknesses may change with as much as 40% over merely 3 years [*Steffensen et al.*, 2008], the assumption of slowly varying layer thicknesses breaks down. To better the performance of the layer detection algorithm over data sections containing fast climate transitions, a more accurate prior might be required. Although layer thicknesses in a subsequent batch are most likely to stay close to the current value – as described by a normal distribution for μ_d of mean 0 and standard deviation 0.1 (figure 6.2.2B) – there is also a small chance that they are changing drastically. A more accurate description for the prior of μ_d would therefore be a weighted sum of two Gaussians distributions: One with a relatively small spread to describe the generally slow changes (this one having the largest weight), and a broad one to increase the probabilities in the tails of the distribution. Imposing such a prior would better allow abrupt changes in layer thicknesses to occur. Yet, in case of a 40% increase in layer thicknesses over just 3 years, this will probably in any case cause problems for the layer detection algorithm.

Annual layer signal parameters

Also the way an annual layer is expressed in the ice core data is climate dependent, and therefore varying with depth. The amount of impurities transported onto the ice sheet and deposited at the drill site (wet deposition versus dry deposition), as well as the seasonality of the deposition events, may depend on the climate regime. In the visual stratigraphy data, the annual layers are generally more pronounced during the cold periods, when more dust is blown onto the ice sheet and deposited as cloudy bands visible in the line-scan data. Hence, an annual layer detecting scheme must be able to allow such changes in layer expression to occur.

It will here be assumed that the changes in annual layer signal with depth can be described solely as a changing mean annual layer signal. Hence, only the parameter describing the mean signal ($\boldsymbol{\varphi}$) is allowed to adapted itself to the data. The remaining parameters (Φ, σ_ε^2) are assumed to stay constant. Depending on the data in question, this may be a very simplistic view: Just as well as changes in climate may affect the mean annual layer signal in the core data, it may as well affect the inter-annual variability of this signal. Likewise, the general white noise level on top of this signal may also be climate dependent. Precisely how much these two parameters change with shifting climate regimes depends on the employed annual layer signal model. Nevertheless, these two parameters are still believed to vary less than the mean annual layer signal, and including them as adaptable parameters would significantly increase the complexity of the algorithm.

6.2.2 Sequential updates of parameters

From the previous section, a simple way of making sequential updates of the parameters used in the layer detection model can be seen. By assuming the priors to be described by very simple probability distributions and tying some of the parameters, the Maximum a Posteriori re-estimation equations derived in section 5.4 can be employed. After each batch, approximate posterior distributions of the adaptable parameters can be utilized as prior for the next.

However, this methodology does not take into account that we do have some prior knowledge on how model parameters depend on each other. Using the approach outlined above, one of the underlying assumptions is that the priors for the individual parameters are independent on each other. In other words, the annual layer thicknesses are allowed to change independently on how the mean annual layer signal is changing with depth. Yet, both annual layer thicknesses and the expression of a layer in the ice core data depend heavily on the climate regime at time of deposition. Assuming their changes to occur independently is therefore not a very good assumption: A climate-induced decrease in layer thicknesses is expected to happen concurrently with an increase in vigor of the cloudy bands in the visual stratigraphy data. Indeed, this will be the case for most of the chemical parameters that can be used for annual layer detection in ice cores, as all of these tend to be more or less strongly depending on climate.

In principle, nothing hinders a prior probability distribution of the parameter set θ, in which the priors of the individual parameter values are correlated. *Chien* [2002] has e.g. developed the resulting update equations for use in speech recognition when assuming the parameters to be linearly dependent. But to take such interdependencies into account

requires a lot of knowledge on how the individual parameters co-varies, something which is difficult to quantify. Furthermore, speaking from a practical point of view, the assumption of covariance of the prior parameters does significantly increase the complexity of the analysis.

On the other hand, a simple form for linear dependence in relation to the fixed parameters, σ_d, Φ, and σ_ε^2, could be introduced very easily. Instead of keeping these parameters constant, they could be allowed to vary with the prior distribution for the remaining parameters. However, the present level of knowledge does not justify a very sophisticated approach on this matter.

When later applying the annual layer detection algorithm to the visual stratigraphy data from NGRIP, only a very simple version of re-iterations are performed. The layer detection algorithm is still under development, and in its present form, Maximum a Posteriori iterations of the parameter values turned out not to be entirely stable.

7. Test of inferred layer boundaries

Having developed a HMM-based layer detection algorithm, a next question arises on how to evaluate the outcome of the algorithm. Its performance will be evaluated using synthetic data (chapter 8) and visual stratigraphy data from NGRIP (chapter 9). For the ice core data, the obtained layer boundaries will be compared to those manually counted in the GICC05 chronology. However, such comparison is not straight-forward, as the manually detected GICC05 layer boundaries themselves are subject to errors as well as uncertainties in depth scale. The following section describes some of the issues to keep in mind when judging the degree of similarity between two layer boundary sequences.

For performance evaluations of the layer detection algorithm, the total number of annual layers within a given depth interval as well as the detailed positions of annual layer boundaries must be considered. Most important for the resulting timescale is a good estimate of the total number of annual layers. However, if judging the algorithm performance only by comparing such numbers, the conclusion may be rather misleading. Given that the layer detection algorithm is endorsed with some information on average layer thicknesses, this knowledge can by itself be exploited to give a decent estimate of the number of annual layers within a given depth interval. This is the case, even if the data series contains no annual layer signal or, equivalently, if the model is not able to detect this signal.

Consequently, the performance of the algorithm should also be assessed based on the similarity between modeled and manually counted layer boundaries. However, for several reasons such comparison of the individual annual layer boundaries is not trivial. For the GICC05 timescale, inaccuracy in depth scale of the high-resolution chemistry measurements (on which the chronology is based) leads to inherent uncertainties in the precise location of designated annual layer boundaries (section 2.3.5). In addition, individual variations in the year-to-year timing of peak concentrations in the various chemical components may cause the designated layer boundaries not to occur simultaneously with e.g. peak values in the visual stratigraphy. For this reason, it was decided to manually transfer the GICC05 annual layer boundaries to the visual stratigraphy data series before comparison.

When matching up the resulting annual layer boundary positions, it must also be kept in mind that neither the GICC05 chronology, nor the Forward-Backward layer detection algorithm developed here, was constructed with the objective of providing an optimal segmentation of the data series into annual layers. Rather, the aim was to obtain a best estimate of the overall layer number. Hence, the comparison takes place between two sets of layer boundaries, none of which are optimal for any such comparison to take place. Besides, one should also keep in mind that the GICC05 chronology is not perfect.

Hence, although the similarity in annual layer boundary positions is an important tool for validating the model results, one should not put too much emphasis on the finer details of the comparison. Certainly, a modeled annual layer boundary should not be dismissed just because its position does not exactly coincide with a layer boundary present in the GICC05 chronology.

7.1 Comparison of layer boundary positions

As previously mentioned, validation of a modeled annual layer sequence must be based on a comparison of layer boundary positions as well as on the overall layer count. The approach followed here has been to separate out the two effects, and evaluate the performance of the algorithm based on a consideration of both of these individually. Such approach is fairly similar to what one might have done per eye if trying to judge the similarity between two sets of layer boundaries.

Differences in layer boundary positions are therefore compared only for layer boundaries which unambiguously can be paired up. The resulting average discrepancy is subsequently compared to that resulting from arbitrarily positioned layer boundaries. Differences in annual layer count are expressed both in the overall number of detected layers, but also by the fraction of annual layer boundaries which could not be paired up.

In the following, the two sets of layer boundaries to be compared are denoted $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{N}$. These $2N$ layer boundaries have been selected from the total number of annual layer boundaries as those whose counterparts in the opposite set are unambiguously defined by a one-to-one mapping: The closest neighbors to x_i and y_i are respectively y_i and x_i. Any linkages involving uncertain layer boundaries in the GICC05 chronology have been omitted from analysis. They are masked out in such a way that they are allowed to be 'closest layers', but they do not contribute to the Δ-value if not being matched up.

To measure the discrepancy between two sets of layer boundary positions coupled by a one-to-one mapping as described above, the average of their squared differences is employed:

$$\Delta^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2$$

In other words, the procedure used for calculating Δ^2 corresponds to a pairing up of the annual layer boundaries in closest pairs, and computing the average squared difference

Box 4: Expectation value of Δ^2 for arbitrary sequence

Consider the layer boundary sequence $\{x_i\}_{i=1}^N$, where the distances between individual boundaries are distributed according to a lognormal distribution. Furthermore, an arbitrary sequence $\{y_i\}_{i=1}^N$ is constructed in such a way that y_i is the point closest to x_i, with equal probabilities of y_i being smaller and larger than x_i. Within each interval, before and after y_i, the distribution is uniform. Denoting by λ_1 and λ_2 the annual layer thicknesses on either side of layer boundary x_i, the probability distribution for a given value $\delta = y_i - x_i$ is:

$$p(\delta|\lambda_1,\lambda_2) = \begin{cases} \frac{1}{\lambda_1}, & -\frac{\lambda_1}{2} \le \delta < 0 \\ \frac{1}{\lambda_2}, & 0 \le \delta \le \frac{\lambda_2}{2} \\ 0, & otherwise \end{cases}$$

The expectation of δ^2 is given by:

$$\begin{aligned}
\mathbb{E}[\delta^2] &= \iiint_{\delta,\lambda_1,\lambda_2}^{\square} \delta^2 p(\delta,\lambda_1,\lambda_2) d\delta d\lambda_1 d\lambda_2 \\
&= \int_{\lambda_1}^{\square}\int_{\delta<0}^{\square} \delta^2 p(\delta|\lambda_1)p(\lambda_1)d\delta\, d\lambda_1 + \int_{\lambda_2}^{\square}\int_{\delta\ge 0}^{\square} \delta^2 p(\delta|\lambda_2)p(\lambda_2)d\delta\, d\lambda_2 \\
&= \int_{\lambda_1=0}^{\lambda_1=\infty}\int_{\delta=-\frac{\lambda_1}{2}}^{\delta=0} \frac{p(\lambda_1)}{\lambda_1}\delta^2 d\delta\, d\lambda_1 + \int_{\lambda_2=0}^{\lambda_2=\infty}\int_{\delta=0}^{\delta=\frac{\lambda_2}{2}} \frac{p(\lambda_2)}{\lambda_2}\delta^2 d\delta\, d\lambda_2 \\
&= \int_{\lambda=0}^{\lambda=\infty}\int_{\delta=-\frac{\lambda}{2}}^{\delta=\frac{\lambda}{2}} \frac{p(\lambda)}{\lambda}\delta^2 d\delta\, d\lambda \\
&= \int_{\lambda=0}^{\lambda=\infty} \frac{p(\lambda)}{\lambda}\left(\int_{\delta=-\frac{\lambda}{2}}^{\delta=\frac{\lambda}{2}} \delta^2 d\delta\right) d\lambda
\end{aligned}$$

The second-last equality can be seen from symmetry considerations. Inserting the probability density function for lognormal distributed layer thicknesses described by parameters μ and σ:

$$p(\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}\lambda}\exp\left(-\frac{(\ln\lambda-\mu)^2}{2\sigma^2}\right)$$

And integrating the above, the following expression is obtained:

$$\begin{aligned}
\mathbb{E}[\delta^2] &= \frac{1}{12}\int_{\lambda=0}^{\lambda=\infty} \frac{\lambda}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\ln\lambda-\mu)^2}{2\sigma^2}\right) d\lambda \\
&= \frac{1}{12}\exp(\mu+\sigma^2)^2\,\mathrm{erfc}\left(\frac{-\mu-2\sigma^2}{\sqrt{2}\sigma}\right) \approx \frac{1}{12}\exp(\mu+\sigma^2)^2
\end{aligned}$$

This is the expected squared discrepancy between a single layer boundary and a neighbouring randomly positioned point. For several layers, the estimated mean of squared differences is exactly the same:

$$\mathbb{E}\left(\Delta^2_{arbitrary}\right) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\delta_i^2\right] = \mathbb{E}[\delta^2] = \frac{1}{12}\exp(\mu+\sigma^2)^2 = \frac{1}{12}\lambda_{eff}^2$$

With an 'effective layer thickness' given by $\lambda_{eff} = \exp(\mu+\sigma^2)$.

between these. This measure of discrepancy, along with the fraction of layers which could not be paired up, is used for evaluating the performance of the layer detection algorithm.

7.1.1 The Δ–value of arbitrary sequences

By the approach just outlined, the layer boundaries to be compared have been selected in such a way that they are as similar to each other as possible. Consequently, even for an arbitrary sequence, the value of Δ^2 will generally be quite small. To evaluate the significance of a calculated Δ^2-value, it must be assessed whether this value is significantly different from that of an arbitrary sequence with an appropriate spacing. Arbitrary sequences with a different spacing can easily be recognized by their disparity in total number of counted annual layers, and by the large fraction of layers not being paired up.

Compare a log-normally distributed layer boundary sequence $\{x_i\}_{i=1}^N$ described by lognormal distribution with parameters μ and σ to an arbitrary sequence $\{y_i\}_{i=1}^N$, with each y_i being the counterpart to x_i. Denoting by λ_1 and λ_2 the annual layer thicknesses before and after x_i, each y_i is constructed to be situated within the range $\left[x_i - \frac{\lambda_1}{2}, x_i + \frac{\lambda_2}{2}\right]$ with equal probabilities of being situated before and after, and with uniform distribution at either side. In this case, it can be shown (see box 4) that the expectation value of Δ^2 is given by:

$$\mathbb{E}\left[\Delta^2_{arbitrary}\right] = \frac{1}{12}\lambda^2_{eff}, \quad \lambda_{eff} = \exp(\mu + \sigma^2)$$

Or equivalently:

$$\mathbb{E}\left[\Delta_{arbitrary}\right] = \frac{1}{\sqrt{12}}\lambda_{eff} \approx 0.29\,\lambda_{eff} \tag{7.1.1}$$

The calculated mean value of $\Delta_{arbitrary}$ is slightly smaller (figure 7.1.1A). The above construction of y_i's does not ensure x_i to also be the layer closest to y_i, which may not be the case if y_i is located near the center of a layer on either side of x_i. Such pairs (x_i, y_i) do not fulfill the requirement of an unambiguous pairing up, and have been removed prior to the calculation of Δ^2. Removal of the above-average contribution from such pairs causes the calculated values of $\Delta_{arbitrary}$ to be slightly less than their theoretical values.

It is not taken into account by the expectation value presented in (4.2.1), that an arbitrary layer sequence obtained by use of the Forward-Backward algorithm is likely itself to be log-normally distributed – even if the algorithm has not managed to properly locate the annual layer boundaries. Such dependency between successive y_i's causes an increased number of the worst aligned y_i's to be discarded due to ambiguous mappings, and hence produces a decrease in the calculated Δ-values (figure 7.1.1C and D).

7.1.2 Evaluating obtained values of the similarity measures

Based on the results of the above investigations (figure 7.1.1), the following conclusions are made: For a sequence of calculated layer boundaries to bear more similarity to a known set of layer boundaries than would an arbitrary sequence, it is required that:

$$\Delta \ll 0.24 \cdot \lambda_{eff}$$

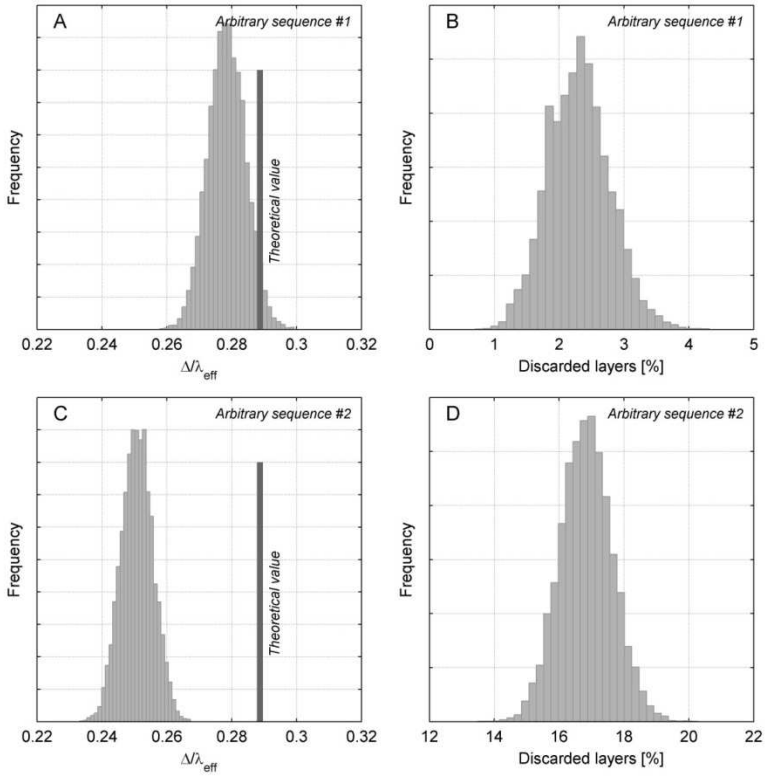


Figure 7.1.1: *A, B: Resulting distribution of Δ/λ_{eff} and percentage of layers discarded prior to analysis, when based on 10.000 realizations of arbitrary sequences constructed as described in section 7.1. Layer thicknesses are assumed log-normally distributed with $\mu = -4.25$ and $\sigma = 0.3$. On average, about 2.5% of the proposed layer boundaries are removed due to ambiguous pairings, causing the calculated Δ-values to be slightly less than those theoretically obtained from equation (4.2.1) (the grey bar). C, D: The analogue results when the arbitrary sequences have a lognormal distributed spacing described by the same parameters as the original layer boundary sequence. In this case, the percentage of removed layers is much larger, and the Δ-value is correspondingly lower.*

This Δ-value should have been obtained without prior removal of too many layers – definitely much less than 14%! – and furthermore, of course, the resulting total number of layers in the data series should be very similar. If the annual layer count in the two data series is very different, the Δ-value does not contain much information.

Mere fulfillment of these three measures does, however, not provide any guarantee that the algorithm works: We wish the algorithm to produce layer boundaries that indeed are far better than an arbitrary sequence! Yet, it must be acknowledged that for the considered depth interval of the NGRIP ice core, the annual layer thicknesses are small. With an average value of λ_{eff} around 1.5 cm, as appropriate for the lower part of the considered depth interval, Δ is required to be much smaller than 3.6 mm.

The developed layer detection algorithm will in chapter 9 be run for the depth interval from 2200 to 2240 m in the NGRIP ice core. For this depth interval, the conductivity profile provided the main support for placing the GICC05 layer boundaries, and annual layer locations were generally placed at peak values in the conductivity profile. To allevi-

ate the problems of different depth scales, different times of peak concentrations etc. between the conductivity and the visual stratigraphy data, the GICC05 layer boundaries for the considered interval were manually transferred from the conductivity data to the visual stratigraphy data.

The fraction of layers which had to be discarded before evaluation of Δ, in the following denoted by F, may also contain information on how well the layering is reproduced. In this measure, layer boundaries removed due to uncertain layers in the GICC05 chronology has not been included. The information in F is in some ways similar to that in Δ: A more precisely reproduced timescale leads to fewer layers which need to be removed. Yet, the information in the two is not completely the same. The value of Δ is sensitive to the small deviations in annual layer positioning, whereas only layers far off any GICC05 counterpart contribute to the value of F. But as only a relatively few layers contribute to F, this measure is not as statistically robust as Δ ,when considering short sections.

The combination of considering the number of counted layers, and evaluating the resulting Δ and F-values can help to assess how well the annual layer detection algorithm performs. However, none of them are a perfect measure, and if the number of layers in the two sequences to be compared are not fairly similar, the information in Δ and F may even be misleading.

8. A sensitivity analysis

In this chapter, the annual layer detection methodology outlined in the previous chapter will be investigated for its resistance to noise in the data series and incomplete knowledge on the appropriate model parameters.

The sensitivity analysis is performed on synthetic data. These present an opportunity to obtain an estimate of the performance of the algorithm in a very simple case, as well as they have the advantage that the outcome of the layer detection algorithm can be compared to a result, which is known with certainty.

8.1 Construction of synthetic data series

Unless otherwise specified, the synthetic data in the subsequent sections have been constructed in the following way: The first layer is assumed to start at the same depth as the first observation in the data series. Annual layers are then produced repeatedly by each time selecting at random the layer duration from a prescribed duration probability distribution. The annual layer thicknesses are taken to be lognormal distributed with the following parameters:

$$\lambda \sim \mathrm{Log}\,\mathcal{N}(-4.25, 0.3^2)$$

These values are chosen as to be similar to those governing the annual layer distribution in the NGRIP ice core for a large part of the considered depth interval.

Having determined the annual layer thicknesses, the annual layer trajectories are chosen based on the probability distributions of the annual layer signal parameters. Finally, Gaussian white noise with variance σ_ε^2 is added to these trajectories, and a synthetic data series has been generated.

To simplify the results, a very simple annual layer model with just a single free parameter has been used: The annual layers are constructed as sinusoidal waveforms of different durations and amplitudes. A sine function has been selected as basis function to ensure the underlying trajectory to be continuous across layer boundaries, and therefore not producing discontinuities that might artificially help the algorithm in its search for annual layers.

The only layer signal parameter is therefore the amplitude, which is distributed according to a Gaussian distribution with mean φ and variance Φ.

The parameter values describing the inter-annual variance in layer shape and white noise component of the visual stratigraphy data is much dependent on the climatic regime, the model employed and the preprocessing of observations taking place prior to analysis. For the sensitivity studies, the value of these two annual parameters are generally set equal to $\Phi = 0.5^2$ and $\sigma_\varepsilon^2 = 0.5^2$. These values present a mean of the range of parameter values investigated.

The length of each observation sequence is chosen such that each sequence on average contains about 50 annual layers, and the statistics are based on an ensemble of 200 realizations of such sequences. Throughout the chapter, the notation $\langle\cdot\rangle$ will be used to denote the mean of the ensemble distributions, and $\mathrm{std}(\cdot)$ will denote their standard deviation.

8.2 Sensitivity to annual layer variability

In this section, the stability of the layer detection algorithm is investigated with regard to how clearly distinguishable the annual layers appear in the data series. It should be no surprise that the annual layers are easiest to identify if the individual layer thicknesses are relatively similar, all layer trajectories are fairly identical, and the additive white noise component is small.

To investigate the importance of each of these three factors for the performance of the layer detection algorithm, the algorithm is first run in the least challenging way: The model parameters are assumed known, and the most likely annual layering is inferred based on this knowledge. Hence, as no EM-iterations need to be performed, the results solely depend on the performance of the Forward-Backward and the Viterbi algorithm respectively.

The analysis is performed for multiple ensembles of synthetic data series, which have been generated with an increasing degree of variance among individual layer shapes and with an increasing degree of additive white noise. Also data series with an increased amount of variance in layer thicknesses are considered. The performance is evaluated based on the resulting counting discrepancy relative to the original data series (ΔN), the average displacement of layer boundaries (Δ), and the percentage of layer boundaries (F) discarded in the calculation of this quantity due to a lacking counterpart in the opposite set of layer boundaries.

8.2.1 Inter-annual variations in layer shape

The performance of the HMM layer detection algorithm will first be investigated for data series with various degrees of inter-annual variability in layer shape. The annual layers are constructed as sinusoidal waveforms of different durations and amplitudes. Their mean amplitude is held constant at 1, while their spread around this value is varied from 0.1 to 1. With a value equal to 1, the spread in amplitude of the respective waveforms is equal to their mean amplitude, and consequently there is a large probability for each layer to have

an amplitude around zero, and there is even 16% chance of the amplitude parameter being negative.

In figure 8.2.1 is shown the performance of the HMM layer detection algorithm for multiple ensembles of such data series, generated with an increasing degree of inter-annual variability in layer shape. For all ensembles shown, the spread of the white noise component is kept constant at $\sigma_\varepsilon^2 = 0.5^2$. Examples of small sections of observation sequences generated based on the given parameters, along with the original and the reconstructed layer boundaries, are displayed in figure 8.2.1A. The sections are taken from the middle of the observation sequence, such that any memory effect due to knowledge of the location of the very first layer boundary essentially is eliminated. Annual layer boundaries are placed according to the output of the Forward-backward algorithm.

The layer detection model is seen to work very robustly, even when the annual layers display a wide range of amplitudes. This is due to the algorithm being able to infer the best layering based on the entire observation sequence. Even if a single layer is buried in noise and indistinguishable from the surroundings, there is a good chance that the surrounding layers are identifiable, and the layer detection model will then try to place a layer boundary at an appropriate place, this being determined based on layer thicknesses. And indeed, with increasing variance of the amplitude of individual layers, not only the percentage of layers with negligible signals of almost zero amplitudes have increased, also increased has the percentage of easily detectable layers having very large amplitudes.

An example of how the algorithm is able to take the annual layer thicknesses into account is found for $\Phi = 1$ between 30 and 35 cm. At 31 cm, a well-defined layer is ending, and the next distinct layer only starts around 36 cm. In between these two layer boundaries, the annual layer signal in the data series is so poorly defined that it basically no longer exists. Yet, even in this difficult case, the knowledge of the two surrounding layer boundaries aids the algorithm to rightfully decide that another two layer boundaries should be located in between these. These layer boundaries may not be placed very accurately, but they are placed, and that is what matters for the resulting timescale.

Total number of counted layers

Regardless of the variance levels of the random component and the white noise component, the difference in number of inferred and original annual layers produces a symmetrical distribution with an average of zero (figure 8.2.1B). This is very fortunate, as it implies that the annual layer detection model is not biased towards either too thick or too thin annual layers.

However, with increased inter-annual variations in layer shape, more counting mistakes occur, and the spread of the ΔN-distribution around zero increases. For the smallest degree of amplitude variations investigated ($\Phi = 0.2^2$), more than 95% are doing a perfect job counting-wise, and are counting exactly the right number of layers. This number is reduced to 80% for $\Phi = 0.4^2$, and is going down to 50% for amplitude variations of the same size as the original underlying signal ($\Phi = 1^2$). Still, this is quite a high percentage. And even in this case, for most realizations the inferred number of annual layers is within ± 1 from the original number of layers. As the entire section includes around 50 layers on average, this gives rise to maximum counting errors around 2%.

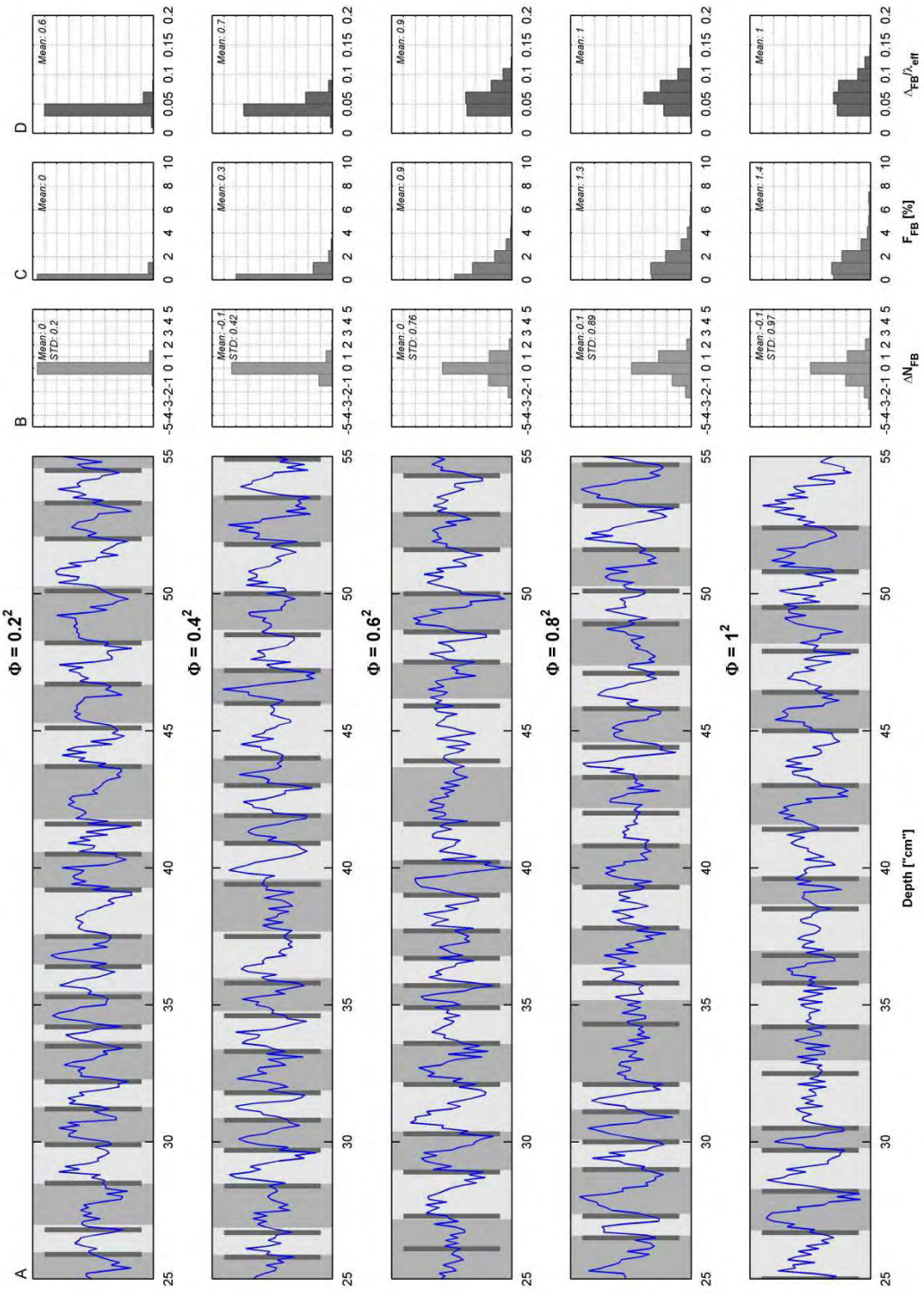


Figure 8.2.1: *Performance of the layer detection algorithm for synthetic data series generated with an increased amount of inter-annual variability in layer shape. A) A section of a random data series and its original annual layering (alternating dark and light grey banding). The inferred layer boundary positions as found by the Forward-Backward algorithm are shown as dark grey bars on top. To the right (B, C, D) is shown some statistics of the performance, based on an ensemble of observation sequences similarly generated. Each horizontal grid-line is 10 percentage points.*

Discarded layers and Δ-value

With increasing degree of variation among the individual layer shapes, the annual layer boundaries become increasingly diffuse and difficult to locate. This implies that even if the total number of counted layers turns out correct, the annual layer boundaries will be placed less precisely. As a result, the calculated Δ-values as well as the percentage of discarded layer boundaries (F_{FB}) steadily increase with enlarged variance in the individual layer shapes. Furthermore, the total number of annual layers may be counted correctly even if there is one layer too much at one location and one too little at another spot, whereas such an event will show up as an increase in either Δ- or F-value. These two measures therefore provide a better measure of how well the layering in details is reproduced.

In case of just small variations in annual layer shapes, almost no layers need to be discarded ($F_{FB} \approx 0$), and they are placed very accurately indeed; the mean of Δ/λ_{eff} is around 0.04 for $\Phi = 0.2^2$. Even when large annual variations in layer shape are allowed, however, the average value of Δ/λ_{eff} is significantly smaller (0.07) than that of an arbitrary sequence (0.24).

8.2.2 The white noise component

Figure 8.2.2 shows the result from multiple ensembles of observation sequences which have been constructed with an increasing level of additive Gaussian white noise. The amplitude variations of the annual layer signal is fixed at $\Phi = 0.5^2$. The variance of the white noise component is increased from 0.2^2 to 1^2, i.e. to the same level as the amplitude of the underlying signal itself.

Difference in layer counts

Again, the annual layer detection model is seen to perform well. The deviations in layer count are symmetrically distributed around 0, and for all values of the white noise variance a major part of the realizations end up with a correct estimate for the number of annual layers in the observation sequence.

For a relatively small value of $\sigma_\varepsilon^2 = 0.2^2$, exactly the right number of annual layers is inferred for more than 90% of all realizations. Of course, the performance degrades with increasing amounts of noise, which shows up as a slow but steady broadening of the histogram showing the discrepancy of annual layer counts. Yet, even when the white noise component is of the same magnitude as the mean amplitude of the signal, the algorithm is counting the right number of layers in 40% of the cases, and almost all realizations are counted within 2 layers hereof. Having on average 50 layers within each observation sequence, this amounts to a maximum counting error of 4%.

Discarded layers and resulting Δ-value

With increased noise levels, both the number of discarded layers, F_{FB}, and the average value of Δ increases. For a small white noise variance equal to $\sigma_\varepsilon^2 = 0.2^2$, the average of F_{FB} is just 0.1. It is increasing up to 2.5% for large white noise levels, i.e. just 2.5% of all layer boundaries in the two sets cannot be paired up. In this case, the Δ-value has increased to 0.1, which is still significantly different from that of an arbitrary sequence.

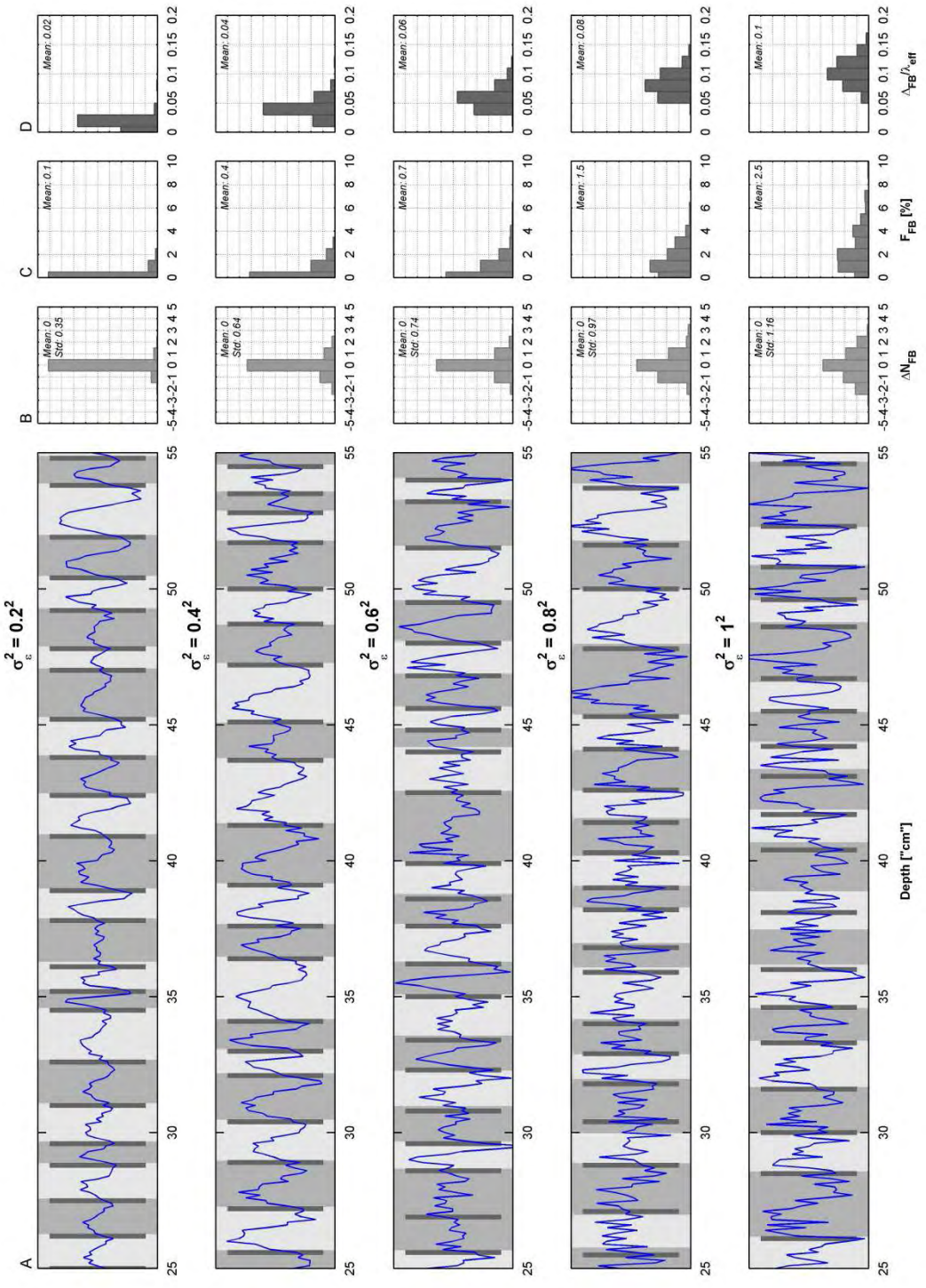


Figure 8.2.2: *Performance of the layer detection algorithm for synthetic data which have been generated with an increased variance of the white noise component. A) A section of a random data series and its original annual layering (alternating dark and light grey banding). Inferred layer boundary positions as found by the Forward-Backward algorithm are shown as dark grey bars. To the right (B, C, D) some statistics of the performance, based on an ensemble of observation sequences, are found. Each horizontal grid-line is 10 percentage points.*

8.2.3 Comparing the two types of layer variability

The annual layering is best reconstructed where the annual layer signal displays the least variation from one year to the next, regardless of whether their differences are caused by varying amplitudes or the addition of white noise. But the two parameters do not influence the performance of the algorithm exactly the same way. To compare their relative importance, ensembles have been generated of all combinations of these two types of annual layer variability, and the performance of the algorithm has been evaluated (figure 8.2.3).

In general, the performance of the layer detection algorithm is most affected by the amount of white noise added to the annual layer trajectories. This is in particular the case when considering the exact placement of the layer boundaries as demonstrated by the changes in Δ and F. For low white noise levels, the amplitude variations of the annual layer signal is almost irrelevant, whereas significantly more degradation results from keeping the amplitude variations small, and increasing the white noise term (figure 8.2.3B,C). Also the resulting discrepancy in number of counted layers is most affected by the addition of large amounts of white noise, although this quantity is affected in a more similar manner by the two types of layer variability (figure 8.2.3A).

The higher sensitivity to the additive white noise component than to the variations in layer shape can be explained as follows: Increasing the amplitude variations does not only produce many layers with a layer expression poorly expressed in the observations – simultaneously, it increases the number of clearly defined annual layers. A high value of the additive white noise component, on the other hand, simply camouflages the annual layer signals everywhere.

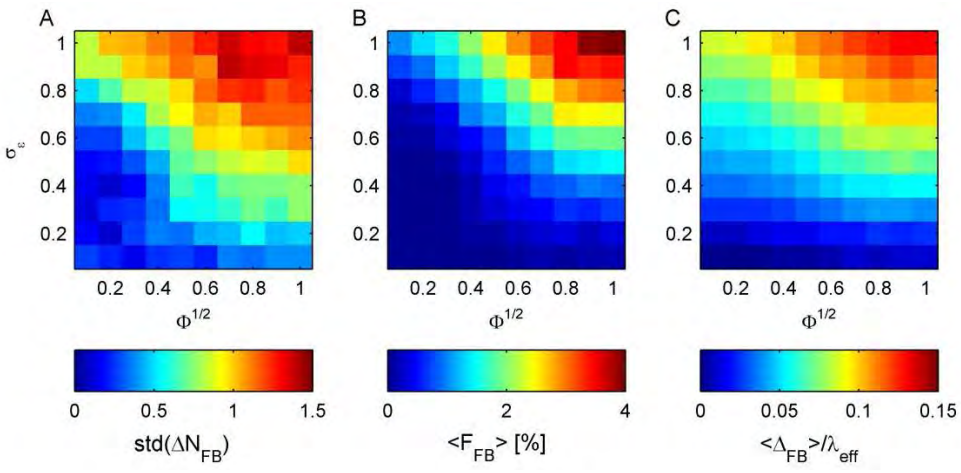


Figure 8.2.3: *Performance of the layer detection algorithm for various values of amplitude variation of the annual layer signal and standard deviation of the Gaussian white noise level. The performance is judged based on standard deviation of the number of miscounted layers (A) and the fraction of discarded layers (B) in the evaluation of the mean deviation of annual layer boundary displacement (C) .*

8.3 Comparison between the Viterbi and Forward-Backward algorithm

For small variance and noise levels, the results based on respectively the Viterbi and the Forward-Backward algorithm are almost identical. For larger variations in annual layer expressions, however, a few differences arise.

The Viterbi algorithm seeks the most likely segmentation of the observation sequence into annual layers, and hence seeks to optimize the positioning of the layer boundaries. This is not the case for the Forward-Backward algorithm. As a result, the Viterbi algorithm generally obtains the smallest values of the layer boundary displacement quantity Δ (figure 8.2.1C).

On the other hand, the Forward-Backward algorithm is supposed to estimate the correct number of annual layers slightly more often than does the Viterbi algorithm, therefore making the Forward-Backward algorithm superior for establishing a timescale of an observation sequence. However, for the chosen model and model parameters applied here, the two methods basically always come up with the same result (figure 8.2.1A), and thus this cannot be confirmed. In any case, however, the Forward-Backward has the major advantage relative to the Viterbi algorithm that it allows an uncertainty estimate on the resulting counting to be evaluated directly.

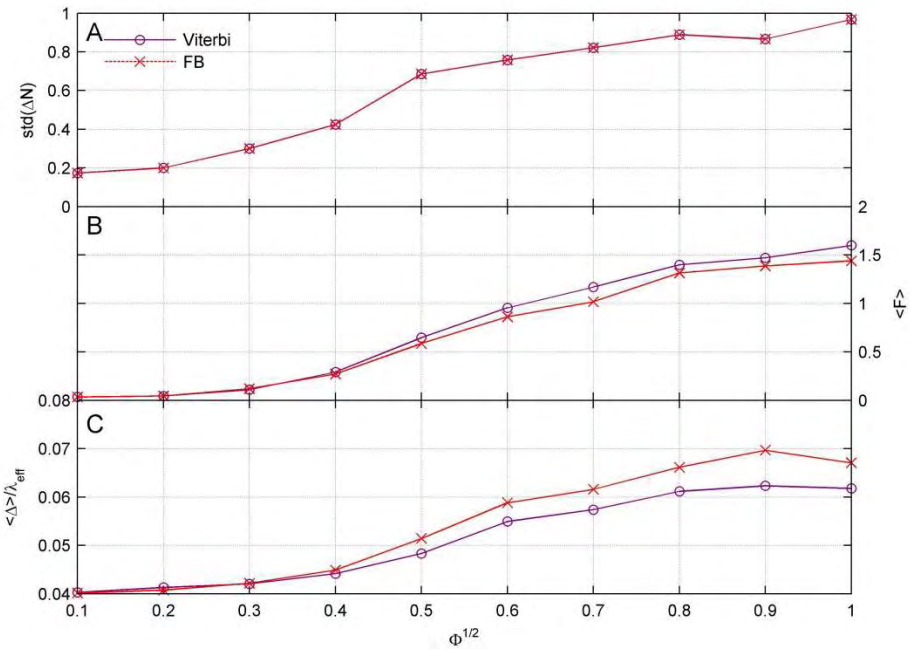


Figure 8.3.1: *Performance of the two layer detection algorithms: The Forward-Backward algorithm and the Viterbi-algorithm. The results are based on an ensemble of runs with parameter values* $\Phi = 0.5^2$ *and* $\sigma_\varepsilon^2 = 0.5^2$.

8.4 Reliability of inferred uncertainty estimates

Inherent to the Forward-Backward algorithm used for annual layer counting is the simultaneous derivation of an uncertainty estimate on the counting accuracy. In this section, the reliability of such uncertainty estimate will be evaluated.

The inferred uncertainty estimate is evaluated by considering each realization in an ensemble at its own, and judging if the original number of annual layers is within the uncertainty bounds of the inferred layer count. The percentage of realizations in which the original number of layers is within the estimated 50% and 95% posterior uncertainty bounds is given in table 8.4.1. For all combinations, the estimated uncertainty bounds are seen to be very reliable: The original number of layers is generally within the estimated 25% and 75% quantiles in ~50% of the cases, and ~95% are within the estimated 2.5% and 97.5% quantiles of the posterior distribution – just as they are supposed to be.

In general, the derived uncertainty estimates even seem to be slightly conservative, which can be explained by the general rounding up of quantile estimates due to the annual layer number being a discrete quantity. This effect is largest when the derived uncertainties are smallest, and hence cannot be expected to hold for real observation sequences.

	$N_{original} \in Q_{50}$ [%]	$N_{original} \in Q_{95}$ [%]
$\Phi = 0.5^2, \sigma_\varepsilon^2 = 0.5^2$	66%	98%
All ensembles	68%	98%
Worst ensemble member	36%	95%

Table 8.4.1: *The reliability of the 50% and 95% uncertainty estimates were estimated for* $\Phi = \{0.1^2, 0.2^2, \ldots, 1\}$ *and* $\sigma_\varepsilon^2 = \{0.1^2, 0.2^2, \ldots, 1\}$. Q_{50} *and* Q_{95} *are the 50% and 95% confidence intervals. The result based on the entire array of ensembles is given. Also given are the results for the member of the 100 ensembles with the lowest percentage of annual layer counts within their allowed range. For* Q_{50}*, this was found for* $(\Phi, \sigma_\varepsilon^2) = (0.9^2, 1)$*, and for* Q_{95} *it was found for* $(\Phi, \sigma_\varepsilon^2) = (1, 0.1^2)$. *The result for the basis ensemble, using* $\Phi = 0.5^2$ *and* $\sigma_\varepsilon^2 = 0.5^2$*, is given as an explicit example.*

However, it must be stressed that the validity of these uncertainty estimates is contingent on the annual layer model and corresponding model parameters to be valid. The derived uncertainty estimate does not include uncertainties contained in these two, and should therefore always be regarded as a lower bound estimate.

8.5 Obtained parameter estimates

To evaluate the bias of the HMM layer detection methodology, the reconstructed parameter estimates for the five model parameters based on the result of the Forward-Backward algorithm have been investigated. For the use of the EM-algorithm, it is important that these parameter estimates are reliable, and that – in case of any bias – this bias is known, such that it can be dealt with appropriately.

These reconstructed parameter values are highly dependent on the inferred segmentation of the data series into annual layers, and hence are also a valuable tool for determining how well the algorithm works under adverse conditions.

In figure 8.5.1, the percentagewise deviation between reconstructed and original parameter estimates for multiple ensembles of observation sequence realizations are plotted. Their respective deviations have been plotted for each of the five model parameters, and for the full range of annual layer shape variations as described by the chosen values of Φ and σ_ε^2.

In general, the parameters describing the appearance of any given layer are more accurately determined than parameters describing the variance between individual layers: The percentagewise deviation of μ_d and $\boldsymbol{\varphi}$ is less than 0.5%, while the remaining parameters are only determined within $\pm$2-5%. This is not unexpected, as it is in general much easier to estimate mean values than standard deviations.

The two parameters describing the annual layer thickness distribution (μ_d and σ_d) are for all ensembles determined within $\pm$0.2% and $\pm$3% of their respective original values. However, while the mean of the layer distribution (μ_d) generally are determined very precisely and seemingly without any bias (figure 8.5.1A: equally many are determined above and below their original values, and no pattern appears to exist), all ensembles tend to underestimate the variance between individual layer thicknesses (figure 8.5.1B). In other word, the layer detection algorithm generally tends to place the layer boundaries a little too regularly. The reason for this probably lies within the way that the algorithm works: Whenever a layer is difficult to place, the layer distribution probability density function is used to derive the most likely layer boundaries. Implicitly, this gives rise to layer boundaries which are a bit too regularly spaced, in particular when the algorithm has to interpolate over longer distances from where the annual layer signal is lost. The result is a layer thickness distribution whose scale parameter is a little too small than what the original data gave rise to.

Although this is undesirable, it is also an unpreventable short-cut that one may have to accept in this kind of analysis. Whenever a layer detection algorithm includes information on the mean annual layer thickness, such information will implicitly give rise to a thickness distribution conforming to this layer thickness. It may therefore eradicate layers which do not seem to agree with this information, in particular if their layer expression is vague. However, it should be noted that although the retrieved spread of the annual layer thickness is biased towards too low values, the retrieved values for this parameter are always less than 3% off from the original. This is negligible, and will not have any perceptible influence on the reconstructed mean annual layer thicknesses.

The mean layer trajectory parameter, $\boldsymbol{\varphi}$, is determined extremely precisely. This is in particular true for small levels of white noise, but even for large values, this parameter is being determined with a precision of $\pm$0.5%. No bias appears to exist. The robustness of this estimate even for large parameter values is due to the ability of the method to place more emphasis on derived parameter estimates from sections of the observation sequence where the layering is the most obvious.

The parameters describing the two types of variation in the annual layer expressions (Φ and σ_ε^2) also seem to be determined without any major bias. However, for small white noise variance values, the discrepancy is largest, and biased towards too high estimates of both Φ and σ_ε^2 . This may partly have to do with the very low values of σ_ε^2 to which they

are compared, making even very small absolute discrepancies show up as large percentagewise deviations.

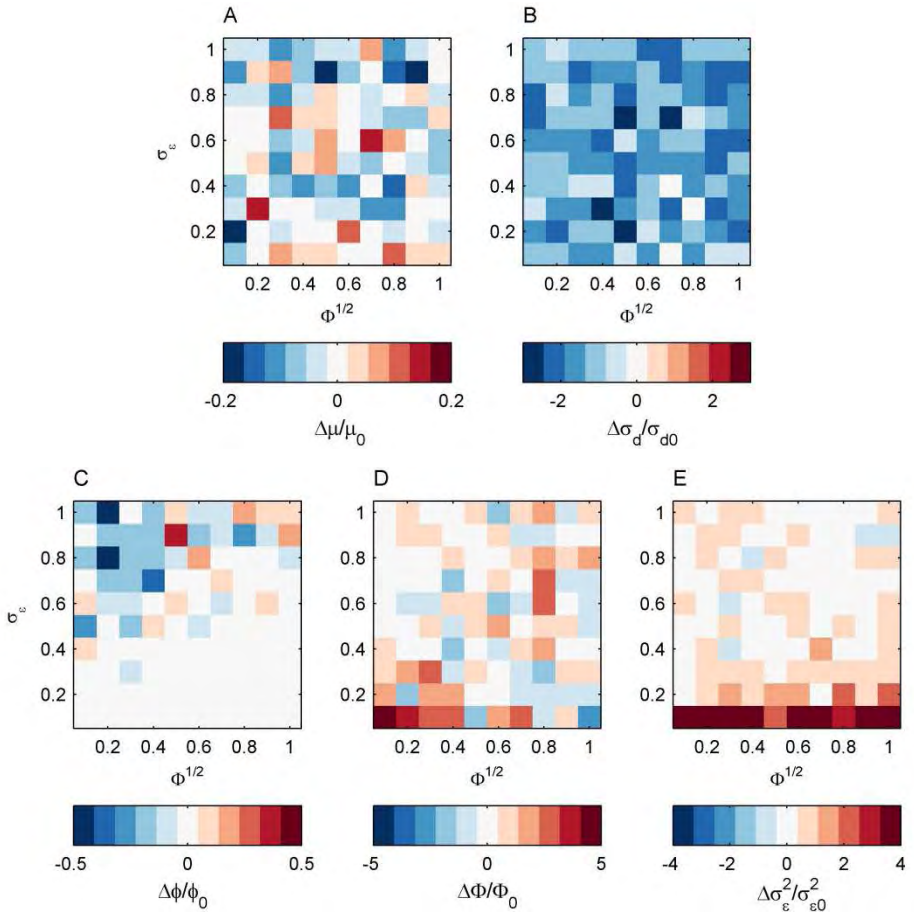


Figure 8.5.1: *Obtained parameter estimates for the Forward-Backward algorithm for different levels of variance and noise in the annual layer signals. The percentagewise discrepancy from their original values (denoted as μ_0 etc.) is plotted.*

8.6 Sensitivity to an erroneous input of model parameter values

The reliability of the inferred annual layering in a batch of data does not only depend on the amount of variability in the data that is not related to the annual signal. It is also dependent on a properly chosen set of parameter values to describe the annual signal as well as its allowed variability. With help of the EM-algorithm it is possible to circumvent the dependence of the chosen initial model parameters. By continuously iterating using the re-estimation equations derived in section 5.3, the most likely set of parameter values for a given data sequence can be found. In this section, it will be investigated how sensitive the layer detection algorithm is to an erroneous input of model parameter values. Does the layer detection algorithm manage to find a correct estimate of these? And how many iterations are required before convergence has been reached?

All re-estimations in the subsequent sections are made using the Maximum Likelihood mode. Apart from an initial (wrong) estimate, the algorithm is not given any information on the original set of layer parameters which have created the data series. Similar sensitivity studies could have been performed in Maximum a Posteriori mode of layer parameter re-estimation, as formulated in section 5.4. However, such studies would not give much additional information on the sensitivity of the algorithm. Contingent on the appropriateness of the prior information used as input, a Maximum a Posteriori approach will simply converge faster.

The rate, with which the algorithm is converging to a Maximum Likelihood estimate of the model parameters, is dependent on the degree of variability among the individual annual layer signals. For all sensitivity studies below, the data series are constructed using the model parameters $\mu_d = -4.25$, $\sigma_d = 0.3$, $\varphi = 1$, $\Phi = 0.5^2$, and $\sigma_\varepsilon^2 = 0.5^2$, which also were used previously.

The sensitivity studies are carried out as follows: One of the five parameters at a time is initiated at a wrong value. The remaining parameters are initiated using their correct value. From here on, all parameter values are allowed to vary as they prefer. The algorithm is asked to perform 10 iterations of re-estimating the parameter values. This is done for an ensemble of 200 different observation series of approximately 50 layers each, which have all been constructed based on the same set of parameter values. At each step of the iterations, the distribution of the individual parameter values is considered based on the ensemble results. These distributions can be summarized using e.g. median and quartiles, and the evolution of these with iteration number portrays the convergence properties of the annual layer detection algorithm.

Even after having reached convergence, the distributions of the annual layer parameters have a certain spread. This is to be expected. Each of the observation sequences contain only about 50 layers, and the computation of a best estimate of the layer parameters based on solely 50 layers cannot be done with high accuracy. Take e.g. the layer thickness parameters. The logarithm to the layer durations follows a normal distribution with mean μ_d and standard deviation σ_d. Based on a sample of $N = 50$ layers, the uncertainty of the mean of the layer thickness distribution can only be estimated with a certainty of [*J R Taylor*, 1997]:

$$\sigma_{\mu_d} = \frac{\sigma_d}{\sqrt{N}}$$

which in this case equals 0.04. The uncertainty in the estimation of the standard deviation of the distribution based on merely 50 samples is even larger. The fractional uncertainty in the estimation of σ_d is given as follows [*J R Taylor*, 1997]:

$$fractional\ uncertainty\ in\ \sigma_d = \frac{1}{\sqrt{2(N-1)}}$$

This implies that in our case, the value of σ_d can only be estimated with an uncertainty of 10%. Likewise for the other two parameters defining the spread of a distribution, $\sqrt{\Phi}$ and σ_ε, which can be estimated with an uncertainty of respectively 10% and 3% (as each

observation sequence contains approximately 700 observations). With the chosen value of Φ, the uncertainty of the value of the parameter $\boldsymbol{\varphi}$ is 0.07 (table 8.6.1).

8.6.1 Layer thickness distribution parameters

First, the impact on the model results of a wrong initial estimate of the layer thickness distribution parameters μ_d and σ_d will be investigated. The value of these wrong parameter estimates are chosen as extreme, and sudden changes of this size are not very likely to occur in reality.

Figure 8.6.1 shows the evolution of the five model parameters when starting out with a wrong initial estimate of μ_d, i.e. with a layer thickness far from the original. With the selected initial values of μ_d, the value of this parameter reaches an almost stable level after just 2 iterations. Also the distributions of the remaining variables have at this point reached an almost constant mean. Yet, not all the variables have reached complete convergence yet. The spread of the distribution of $\boldsymbol{\varphi}$ is still increasing, and has not yet reached its theoretical value (table 8.6.1). The standard deviation of the remaining parameters is very close to that theoretically predicted.

The recovered parameters are very close to the original ones. Yet, it should be mentioned that the value of σ_d generally tends to find a level slightly too low. This was also the conclusion from the previous section, where the model parameters used as input to the algorithm were known, and the explanation is the same: It is a consequence of the way

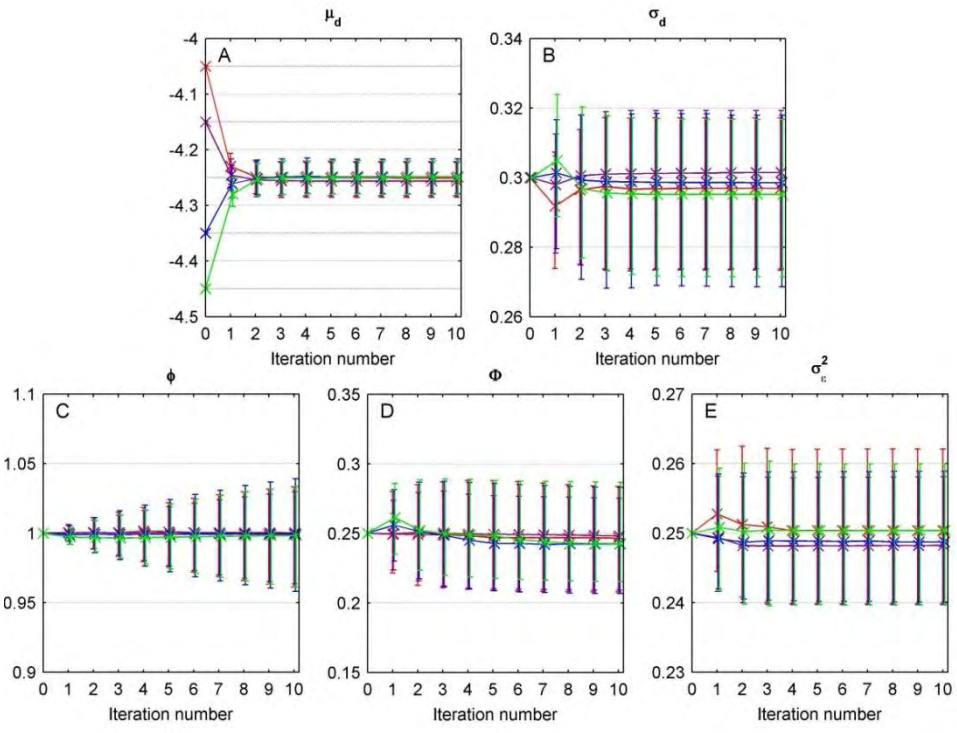


Figure 8.6.1: *Ring-down of parameter values when starting the layer detection algorithm at wrong values of μ_d. The median as well as the 25% and 75% quantiles of the distributions are marked. The original value of μ_d is rediscovered after 2 iterations, and after 10 iterations, only the layer shape parameter $\boldsymbol{\varphi}$ has not yet recovered a completely stable level.*

	μ_d	σ_d	$\boldsymbol{\varphi}$	$\sqrt{\Phi}$	σ_ε
Theoretical mean	-4.25	0.3	1	0.5	0.5
Observed mean	-4.251	0.296	0.998	0.494	0.500
Theoretical STD	0.042	0.030	0.071	0.051	0.0134
Observed STD	0.045	0.034	0.053	0.057	0.0146

Table 8.6.1: *Descriptive statistics (mean and standard deviation (STD)) for the theoretical and observed distribution of the five model parameters. The observed distribution is a composite of those reached after the 10'th iteration, when starting out with wrong estimates of the parameter* μ_d.

that the algorithm interpolates over sections with little layer signal in an attempt to make the best use of these. The same tendency of an inferred value of σ_d, which is slightly too low, applies to all iterations for all parameters.

Also Φ appears to have a slight tendency toward a value too small. However, this might just be a consequence of the parameter $\boldsymbol{\varphi}$ not yet having reached a stable level.

Very much the same scenario unfolds itself for the scale parameter of the layer thickness distribution (figure 8.6.2). With the chosen initial values of σ_d, it here takes the algorithm about 3 iterations to reach stability, and again it reaches stability at a level slightly too low. The parameter $\boldsymbol{\varphi}$ is again the slowest one to converge, and although the mean of the distribution stays around 1, as it should, the standard deviation has not ceased to increase after the 10 iterations.

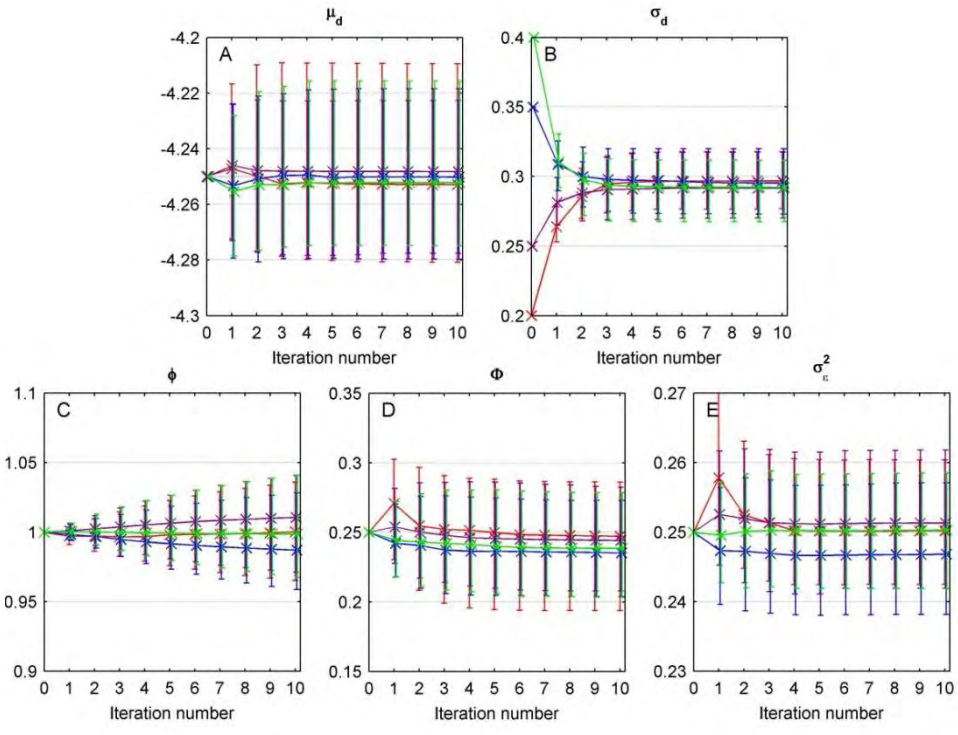


Figure 8.6.2: *Ring-down of parameter values when starting the layer detection at wrong values of* σ_d. *The median as well as the 25% and 75% quantiles of the distributions are marked. The original value of* σ_d *is rediscovered after 3 iterations.*

The parameters describing the distribution of the annual layer thicknesses have a large impact on the resulting number of counted annual layers. Fortunately, it was seen that in this relatively simple case, the model is able to retrieve the correct value of these parameters after just a few iterations of the Forward-Backward algorithm. However, the dependency of a wrong initial input to the model is highly dependent on the overall difficulty of recognizing the annual layers in the data. In case of easy recognizable layers, the large amount of information in the data ensures the algorithm to very quickly converge to an appropriate set of parameter values. In case of less optimal conditions where the annual layering is less apparent, the annual layer detection algorithm is being guided less by the data and consequently more by the annual layer thickness distribution. In this case, a poor initial estimate of the parameters describing this distribution will have larger effect.

8.6.2 Annual signal parameters

The parameter $\boldsymbol{\varphi}$, which describes the mean annual layer signal, turns out to be the parameter which is struggling the most to reach convergence. Even after 10 iterations, none of the runs with different choice of initials parameters has managed to reach a stable level (figure 8.6.3). All of them are heading in the right direction, but in comparison to what was the case for the remaining parameters, convergence is slow.

The slow convergence of $\boldsymbol{\varphi}$ also affects the derived values of Φ. Whereas $\boldsymbol{\varphi}$ describes the mean trajectory of a layer signal, the value of Φ describes the variability of the individual

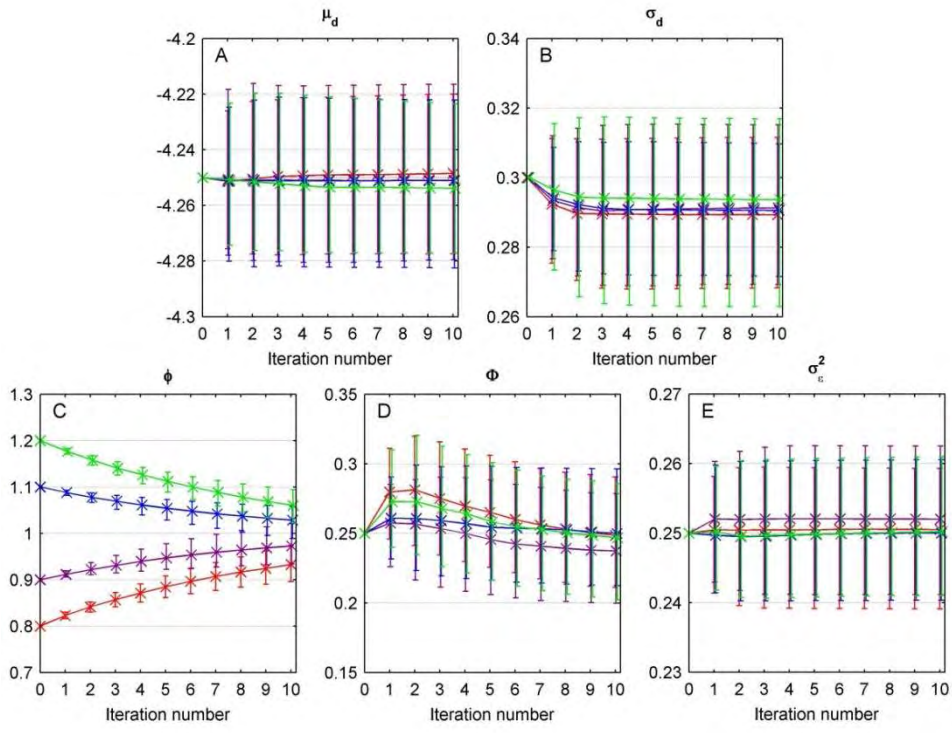


Figure 8.6.3: *Ring-down of parameter values when starting the layer detection at wrong values of the mean annual layer signal parameter* $\boldsymbol{\varphi}$*. The median as well as the 25% and 75% quantiles of the distributions are marked. The value of* $\boldsymbol{\varphi}$ *is seen to converge relatively slowly, and have not reached a stable level after 10 iterations.*

layers around this mean signal. The two are therefore closely connected: The further $\boldsymbol{\varphi}$ is away from its original value, the more variability of the layers around their hereby assumed mean trajectories is required.

The evolution of the annual layer signal parameters Φ and σ_ε^2 (figure 8.6.4 and figure 8.6.5) is very similar to the rest. After three iterations, all parameters have reached a more or less stable level, only the distribution of $\boldsymbol{\varphi}$ may not have reached convergence quite yet. Observe the high accuracy with which the parameter σ_ε^2 is determined. This is due to each observation in the observation sequence contributing to the estimate of this parameter.

To conclude: The annual layer parameters generally behave nice, and are converging towards their Maximum Likelihood estimates, which are almost perfectly the same as the original parameters which went into the construction of the observation sequence to begin with. Only two issues need to be kept in mind: The most critical is that the parameter describing the mean annual layer signal, $\boldsymbol{\varphi}$, is the one which has the most trouble to reach convergence. It does converge, but convergence happens slowly. Hence, the algorithm is the most sensitive to changes in this parameter. The second issue is less of a concern: The value of σ_d is continually being estimated slightly too low. However, as the algorithm relatively easily adjusts the value of σ_d within a few iterations, this should not have much effect on the performance of the layer detection algorithm in practice.

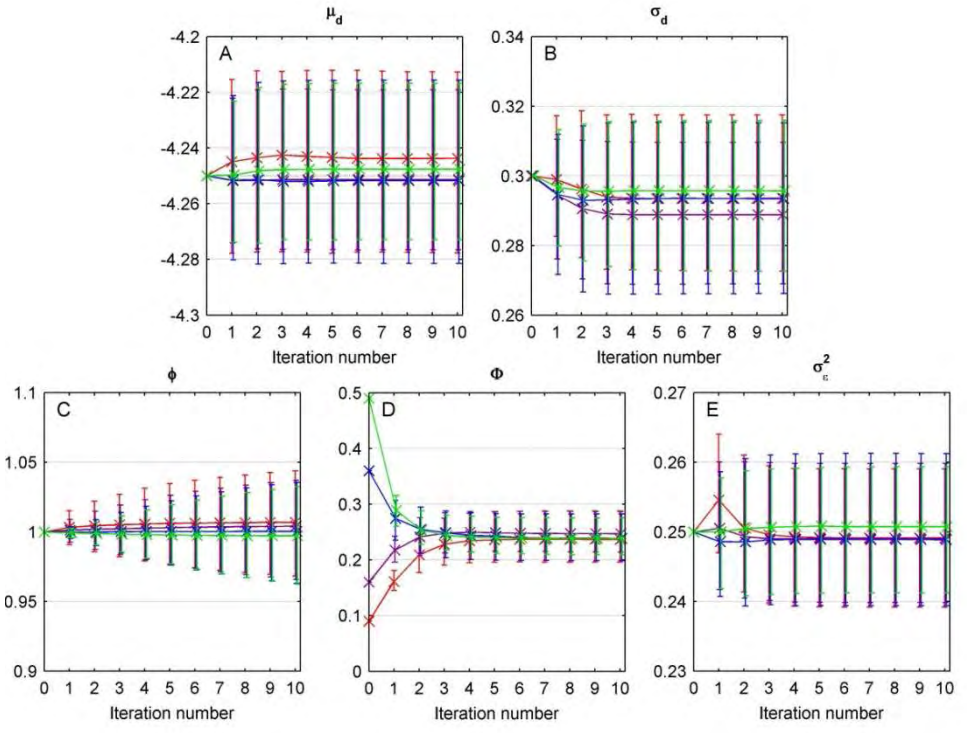


Figure 8.6.4: *Ring-down of parameter values when starting the layer detection at wrong values of Φ, which express the variability of the individual layers around their mean trajectories. The median as well as the 25% and 75% quantiles of the distributions are marked.*

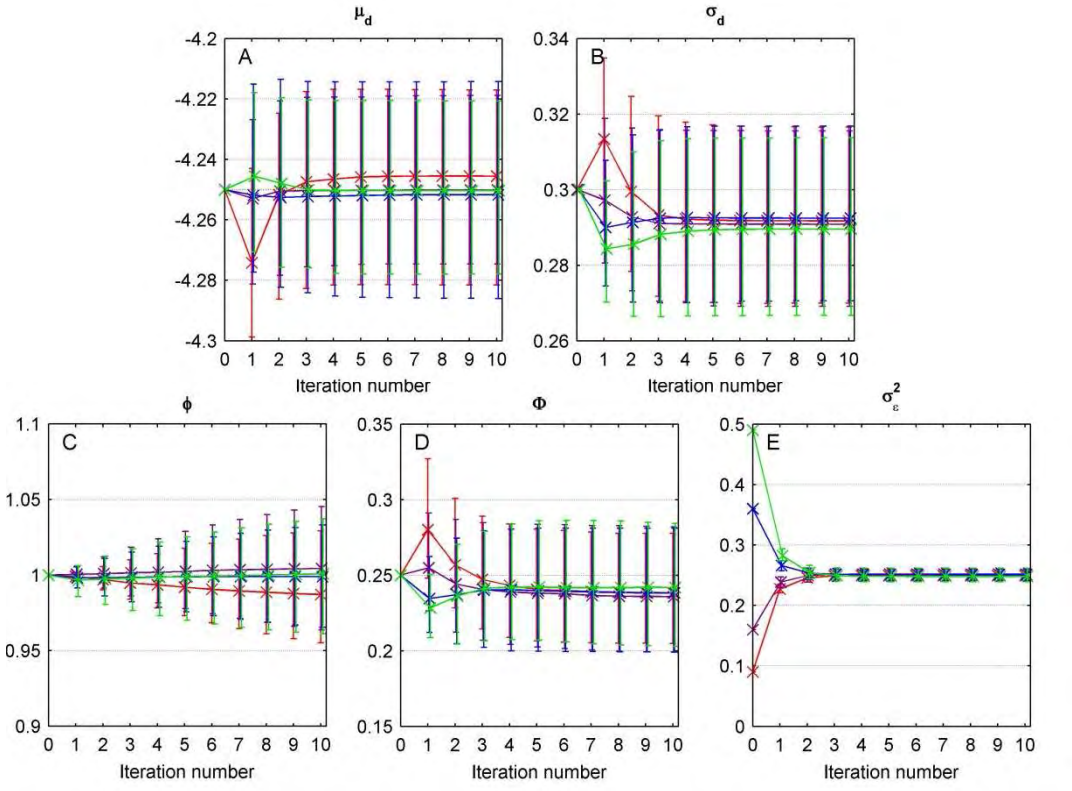


Figure 8.6.5: *Ring-down of parameter values when starting the layer detection at wrong values of σ_ε^2, the variance of the white noise component in the data. The median as well as the 25% and 75% quantiles of the distributions are marked.*

8.6.3 Varying all parameters at once

In the previous section, each of the layer parameters was given a badly chosen initial value, while the rest of the parameters were initiated at their original values. To fully ensure that the convergence of the layer detection algorithm does not depend on the initial set of parameters, a final test was made in which all initial parameter values were chosen freely.

To nonetheless guide the layer detection algorithm a little bit, the initial parameters were randomly chosen as follows: $\mu_d \sim \mathcal{N}(-4.25, 0.2^2)$, $\sigma_d \sim \mathcal{N}(0.3, 0.1^2)$, $\boldsymbol{\varphi} \sim \mathcal{N}(1, 0.1^2)$, $\Phi \sim \mathcal{N}(0.5^2, 0.1^2)$, and $\sigma_\varepsilon^2 \sim \mathcal{N}(0.5^2, 0.1^2)$. The layer detection algorithm was then left to iterate. An ensemble of 500 different observation sequences and differently selected initial parameters was hereby constructed. The resulting distributions of the parameter values after the 10th iteration are shown in figure 8.6.6. Once again, it can be seen that the algorithm has managed to locate the Maximum Likelihood solution of the set of model parameters, which is almost equal to the set that the observation sequences were constructed with. Only the mean layer trajectory parameter, $\boldsymbol{\varphi}$, is struggling. Based on the previous considerations, this is not a surprise. Further investigations indicated that when the mean layer trajectory parameter was chosen very broadly, convergence had simply not yet been achieved after just 10 iterations.

The annual layer model used here is of course a very simple model, and the observations are constructed such as to accurately be described by the model. In reality, things get more complicated. This may result in a decrease of the ability of the iterations to swiftly converge towards a proper set of Maximum Likelihood parameter values. Yet, the relative ease with which the algorithm is able to properly locate the correct set of model parameters in this less demanding case is encouraging.

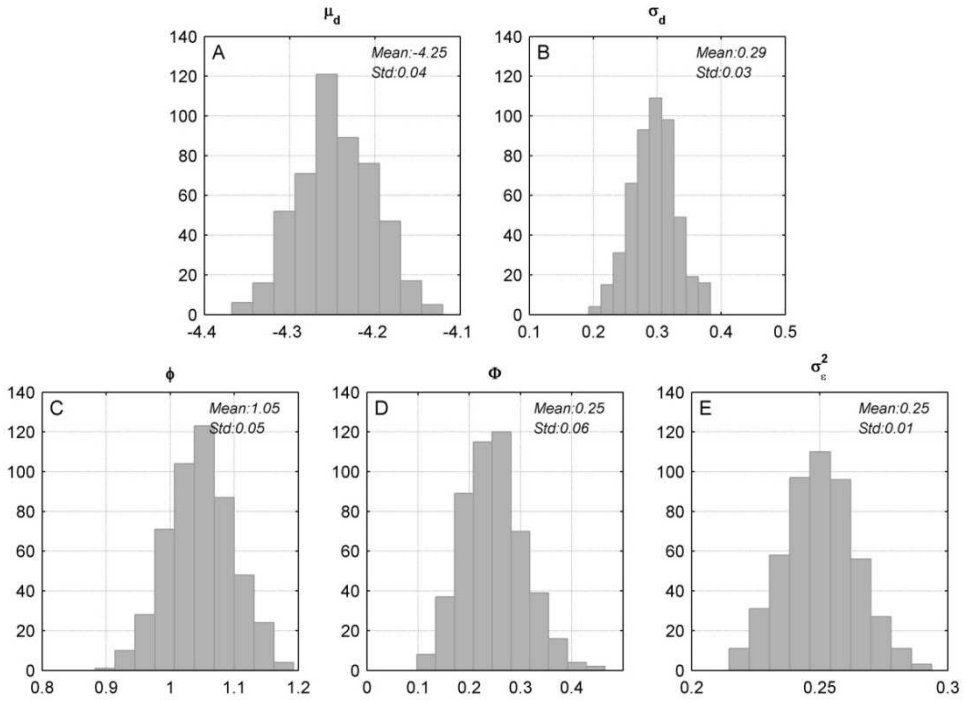


Figure 8.6.6: *Resulting distributions of the inferred parameter values after 10 iterations, when starting from an initial set of parameters randomly selected as described in section 8.6.3.*

8.7 Performance of layer detection algorithm

In general, the annual layer detection algorithm is very robust towards even large degrees of noise in the data. For some of the combinations of parameters with high noise levels, the annual layer signal in the resulting data series would have been difficult to spot with any great deal of certainty by a human investigator. The Viterbi and Forward-Backward algorithms, on the contrary, are able to do this counting in an unbiased manner due to knowledge on the annual layer thickness distribution in combination with a robust measure for judging what is an annual layer.

If provided by an initial wrong input of the model parameters, the algorithm is able to iterate its way to the maximum likelihood parameters. With the employed model, and the chosen noise value of the data, only a few iterations were usually required before convergence was reached. The parameter that converged at the slowest rate was the parameter

describing the mean layer trajectory. It did not manage to reach equilibrium after 10 iterations.

However, the method does have a tendency to obtain too small values for the layer thickness distribution scale parameter, σ_d. The bias of this parameter is a generic imperfection of the algorithm, which simply derives from the desire to use the thickness distribution of the annual layers to determine the most likely layer boundaries in sections where these are not obvious. As a result, the layer distribution will be slightly less broad than it ought to be.

Results here have generally been based on the results of the Forward-Backward algorithm. The performance of the Forward-Backward algorithm and Viterbi algorithm is, however, very similar. They each have their own stronger sides, but in the wish of obtaining an uncertainty estimate associated with the resulting layer counted timescale, the Forward-Backward algorithm is to be preferred. In the following, results based on the Viterbi algorithm will therefore only be used to see how well the two agree with each other, and therefore how much the obtained layering depends on the specific assumptions in the definition of a best timescale for an observation segment.

9. Layer counting in data from different climate regimes

In this chapter, the developed layer detection model based on Hidden Markov Modeling will be applied to the visual stratigraphy data from the NGRIP ice core over a Dansgaard-Oeschger event. Two intervals will be considered: The Greenland Stadial 13 (GS-13, GICC05: 47051-48261 years BP), a cold period from the depth interval 2225-2240m, and the subsequent warm period, the Greenland Interstadial 12 (GI-12, GICC05: 45927-46770 years BP) from the depth interval 2000-2220m. Given that the annual layer signal in the visual stratigraphy data is more distinct during the cold periods, the method is likely to work better here. The layer detection algorithm will also be applied to data from the transitional period between the two climate regimes.

The inferred layering is not to be considered a 'perfect chronology'. The implementation of the layer detection algorithm with regard to the description of an annual layer is still very simple, and compared to the complexity of the annual layer signal in the visual stratigraphy probably too simple. The complexity can be seen from the different outcomes of the layer detection algorithm when using different models for describing the annual layer signal. However, the dissimilarity between individual model results should be seen in context of the visual stratigraphy being a very noisy data sequence. For instance, many annual layers display more than just a single peak. Judged from a visual inspection, the outcome of every one of the models is indeed a probable result. The main difference between the individual model results lies in the differences of judging how many of the extra peaks should be considered to be annual layers.

One way to think of the algorithm with various annual layer models as input is to consider each of these as equivalent to an 'annual layer counter'. Just as every human investigator counting the layering in these data would come up with his/her own result (and different uncertainties), so do these. Some are probably better able to judge from the shape of a peak whether or not it should be counted as a layer, but it is not always immediately clear who is able to do the least biased overall counting. The same is the case here. Yet, in comparison, the force of this algorithm lies in its objectivity of judging an annual layer signal.

Before applying the model to the visual stratigraphy data, the data has been normalized to make the annual layer signal stand out more clearly. This is yet another potential reason for variability between individual model results. Depending on the applied normalization routine, some peaks may be enhanced while others will be degraded, with the overall result being a difference in the resulting number of counted layers. Throughout the results here, the normalization procedure has been kept simple and the same.

The GICC05 chronology in this section of the ice core has mainly been based on the conductivity and visual stratigraphy. For the exact placement of the layer boundaries, the peaks in the conductivity have been used. The conductivity profile was found to show sign of most of the annual layers [*Svensson, pers. comm.*], and its smoothness made it easier to use this profile for annual layer counting. An obstacle in the performance assessment of the annual layer detection algorithm on real data is the existence of small differences in depth scale between the visual stratigraphy and the conductivity data. To alleviate this obstacle, the GICC05 annual layer boundaries have been transferred to the visual stratigraphy before comparison.

The evaluation of the algorithm performance is furthermore made difficult by the GICC05 chronology not necessarily being very accurate when dealing with short intervals [*Svensson et al.*, 2008]. Indeed, the data on which the counting has been based within the considered depth interval does not allow for very firm annual layer detection to be carried out. The annual layers are thin, and only few data series are able to resolve them. However, this is the case for the most part of the depth interval for which high quality visual stratigraphy data exist.

9.1 Preprocessing of data

9.1.1 Normalization of data

Before applying the layer detection routine to the visual stratigraphy data, the data was preprocessed using a normalization routine. This was primarily done with the purpose of increasing the similarity between individual layers. The use of appropriate data preprocessing generally tends to have large influence on the performance of machine learning techniques [*Bishop*, 2006], and this is also the case here.

The general idea behind preprocessing the data is to remove some of background variability not related to the target signal, which in this case is the seasonality signal. For this application, preprocessing was also used to stabilize the heights of the annual layer peaks, thereby removing some of the variability associated with the seasonal signal itself. By reducing the variability of the seasonal pattern, the general task of pattern detection becomes much easier.

Preprocessing of the visual stratigraphy data series can be done in a variety of ways, and only a very simple transformation has been used here. Firstly, data was log-transformed in order to minimize the peak heights of large peaks. Secondly, to remove a varying background signal, the data was then treated by subtracting a running mean. The window

length used for the running mean was 10 cm, and hence contained several annual layers. An example of the resulting data series can be seen in figure 9.1.1.

Other more sophisticated preprocessing procedures of the data series may make the annual layer signal stand out better. Several of these were tested (using Box-Cox transforms [*Madsen*, 2008], normalizing to a constant standard deviation etc.). However, there was no time to investigate these in details, and at this point, it is difficult to justify more complicated procedures. However, it must be stressed that the choice of preprocessing regulates which peaks signals are enhanced and which are not, with large implications for the sensitivity of the algorithm to different types of peaks. In the future, much more time should be devoted to such investigations. Meanwhile, the results produced with the current, very simple, configuration are quite promising.

In figure 9.1.1, the differences in outcome of the algorithm when using the original data (A, B) and the processed data (C, D) are shown. The layer detection algorithm based on the original data shows some skill, but an improvement in performance is evident when including the preprocessing. In both cases, the derived annual layer positions are fairly similar, and the Forward-Backward algorithm ends up with the same most likely estimate of the number of annual layers contained in the sequence (which is one less than in GICC05). Yet, the improvement in performance is indicated by the original data leading to much broader uncertainty bands, and from the difference between the inferred layering when using the Forward-Backward algorithm and the Viterbi algorithm. For the processed data, there is no difference between the two.

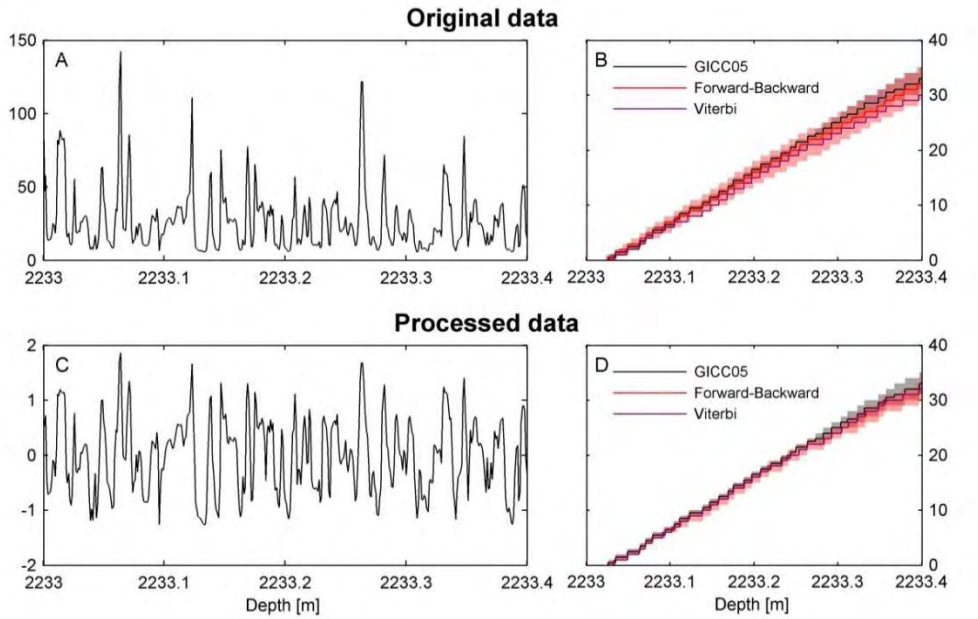


Figure 9.1.1: *Examples of original (A) and preprocessed data (C), and output of the layer detection algorithm using these (B,D). A cosine is used as trajectory function.*

Furthermore, the line-scan data was down-sampled to 1mm resolution. While decreasing the resolution of the data series, such down-sampling decreases the length of each observation sequence drastically. This makes the algorithm run much faster, which again allows for more tests to be carried out. However, the down-sampling does put a limit to the annual layer thicknesses which can be resolved. In practice, this is probably not a problem for the considered data intervals, where annual layer thicknesses are above 1 cm, but it may be a problem for ice core sections where the annual layers are much thinner.

9.1.2 Calculating slope and curvature

Including more information in the annual layer detection algorithm by considering several data series at once should improve the performance of the algorithm. This turned out to be true even if the involved data series were not independent, but derived as derivatives of a master data series.

By taking the derivative of a data series, trends in the data are removed. However, at the same time, the noise-level is enhanced, and the signal-to-noise ratio therefore decreased. A signal in the derivative data series is therefore generally simpler, but noisier. To ensure that the signal in the data series derivative is not completely masked by noise, the visual stratigraphy data series was smoothed prior to taking the derivatives. This was achieved by Savitzky-Golay smoothing, which is a smoothing filter specifically developed for computing the derivative of a noisy data series.

The basic idea behind this type of smoothing is simple. Instead of computing the derivative of a data series by point to point differences, a linear function is fitted to the data in the neighborhood of any given point by linear regression. The slope of the best fitting straight line gives an estimate of the slope of the data here. Likewise, a second order polynomial may be fitted to the data in the surrounding neighborhood, and the derived slope and curvature gives an estimate of the slope and curvature of the data. The polynomial order employed and the size of neighborhood used for fitting determines the degree of smoothing of the data series. In our case, the smoothing was kept to a minimum. Only the immediately surrounding observations were used as neighborhood, a first order polynomial was used to derive the slope of the data series, and a second order polynomial to derive its curvature. The master data series itself was not smoothed.

9.2 Layer detection during a cold period

In this section, the layer detection algorithm will be applied to the visual stratigraphy data from the Greenland Stadial 13 (GS-13). This section covers the depth interval from 2225-2240 m, and a time period of approximately 1200 years (47,051-48,259 years BP according to GICC05). Within this section, the GICC05 chronology predicts a general thinning of layers with depth. This thinning seem to have a counterpart in the δ^{18}O-record, and it is therefore likely to be a real feature (figure 9.2.1).

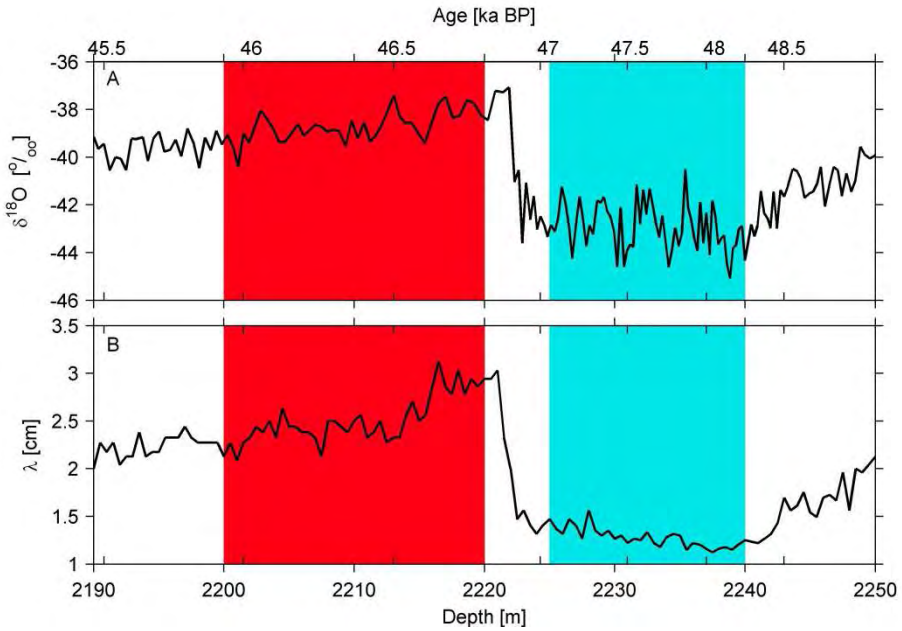


Figure 9.2.1: *Evolution of $\delta^{18}O$ and mean annual layer thickness (λ) over the selected depth interval. The two curves have been treated slightly differently, and are therefore not directly comparable: $\delta^{18}O$ is given as 20 year means, whereas λ-values are given as 50 cm means. The red area is the section considered from the warm period GI-12 (section 9.3). The blue area is the part of the previous stadial, GS-13, which is considered in section 9.2.*

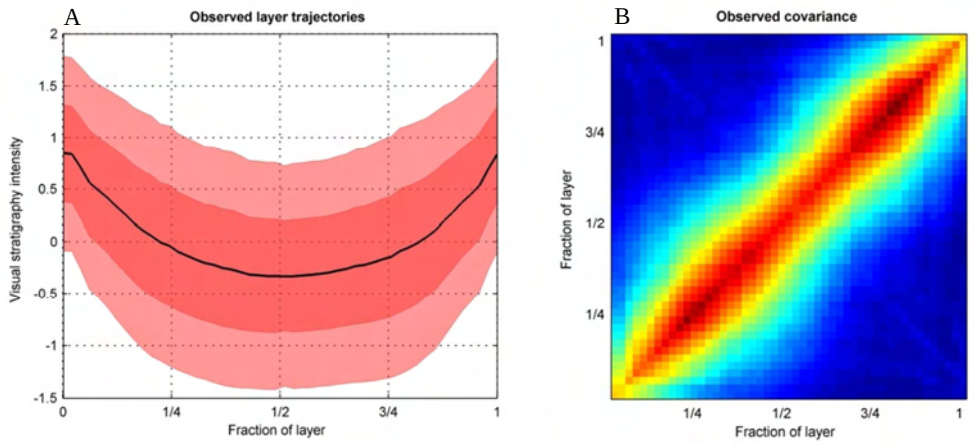


Figure 9.2.2: *The annual layer expression in the visual stratigraphy data during the cold period. A: The mean annual layer shape, and corresponding 1σ and 2σ variance intervals. B: The covariance between individual observations in a layer trajectory.*

Within the interval, climate conditions appear to have been relatively stable, and it is therefore possible to run the algorithm in the simplest way possible: The appropriate parameter values were estimated based on the layers contained within the first 0.5 m of the data, and these parameter values were maintained as fixed values throughout the entire period. Observe that even though the employed parameter values in this way are based on the first part of the data sequence, this does not necessarily imply a flawless performance of the algorithm within this interval.

A range of annual layer signal models were tested. When considering how appropriate these models are, their postulated mean trajectory and covariance between individual observations in a layer should be compared to figure 9.2.2A&B, which shows the trajectory and covariance from the GICC05 layers. The results of the respective annual layer models will be described in the following.

9.2.1 A simple cosine as trajectory function

In figure 9.2.3 the resulting annual layering in the depth interval under consideration is shown, when using a cosine as shape for the layer trajectories (d is duration of the layer):

$$f(x) = A\cos\left(\frac{2\pi x}{d}\right), \quad x \in [0,1]$$

With only a single parameter to adjust, namely the amplitude of the waveform (A), this is one of the simplest functions imaginable which may be able to provide a basis for the annual layer signal. In contrary to the sine function used in the sensitivity studies, fitting a cosine function to the individual layers separately does not imply the joined fitted curve to be continuous. However, by choosing a cosine instead of the sine, the annual layer boundaries will be placed on the peaks, hence allowing a direct comparison to the GICC05 layer boundary positions.

However, from the results it can be seen that the simplicity of the layer shape comes at cost of the performance of the algorithm: Significantly fewer layers are being counted. The discrepancy between the inferred number of layers and the number of layers in GICC05 is 12.6% (figure 9.2.3C). By considering the evolution of the mean layer thicknesses (figure 9.2.3D), it can be seen that the algorithm throughout the section has a tendency of counting too few layers, but lacking most in the last part of the section.

Although the resulting timescale does not agree that well with the GICC05 timescale, most of the inferred annual layer boundaries seem to fit well with GICC05 layer boundaries. A larger example of the inferred layering is included in figure 9.2.7. The computed Δ-value equals 0.21 cm, and using an effective layer thickness of 1.37 cm (an average value found using the GICC05 layer boundaries for the interval), we find that $\Delta/\lambda_{eff} = 0.153$, which is much lower than what was found for an arbitrary sequence (0.24). However, the difference in total layer count between GICC05 and the here employed model may be too large for the similarity measure Δ to be trustworthy.

When looking at a section of counted layers (see e.g. figure 9.2.3E), the reason why too few layers are counted can be found: By defining layer trajectories to be described by a cosine, it is implicitly assumed that two consecutive peaks have the same height. By comparing to the GICC05 annual layers, this is seen to not be a very good assumption. Individual annual peaks do not have similar heights, and hence two consecutive peak values may be very different. As a consequence of not allowing peak values to differ, an annual layer with very dissimilar peak heights at its boundaries is not counted as such. Perhaps the clearest example within the illustrated section is found around 2233.3 m. Here, the algorithm infers the existence of an annual layer which indeed is very broad (and for this reason not very likely), but this is being compensated for by its resemblance to a

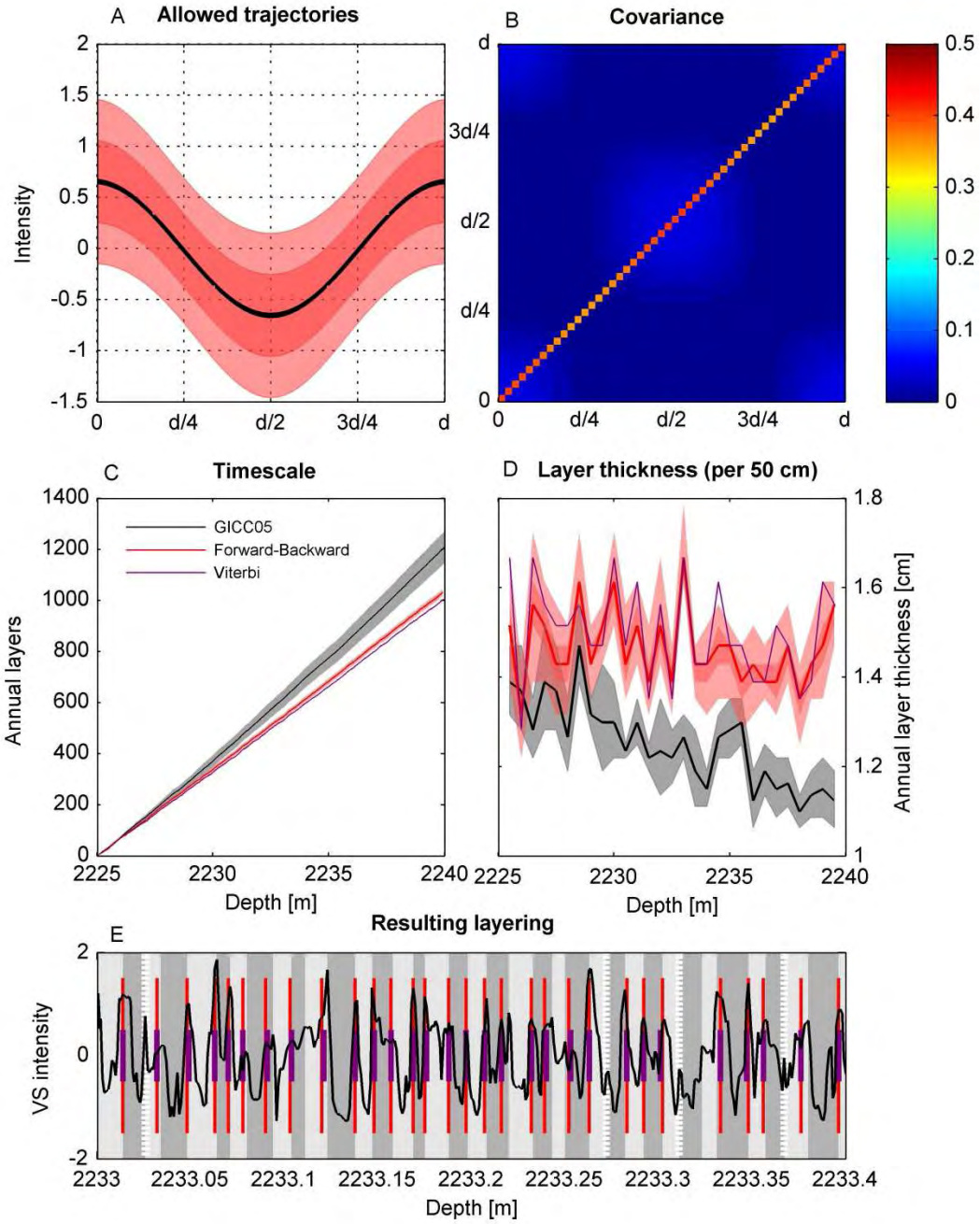


Figure 9.2.3: *Performance of the annual layer detection algorithm when using a cosine as trajectory function. A: Assumed mean layer signal and corresponding 1σ and 2σ-variability bands. B: Assumed covariance matrix for the layer trajectories. C: Resulting inferred timescale when compared to the GICC05. Gray bands show the Maximum Counting Error (MCE) for the GICC05 chronology. Red bands show the 50% and 95% confidence interval for the inferred layering. D: Resulting mean annual layer thicknesses per 50 cm, compared to GICC05. E: An example of the inferred layering. The alternating background pattern of light and dark gray bands show the positions of GICC05 layers boundaries. Uncertain layer boundaries are marked with white horizontal stripes. Red bars are inferred layer boundary positions when using the Forward-Backward algorithm, purple bars show the layer boundary positions based on the Viterbi algorithm.*

cosine. If the section was to be divided up into 2 or 3 layers, as implied by the GICC05 timescale, none of these would be very well described by a cosine.

Indeed, the assumption of a cosine trajectory does not describe the mean annual layer trajectory very well (compare figure 9.2.3A to figure 9.2.2A), and nor does it pick up much of the real variability of the annual layers. This can be seen from the postulated covariance of the layer trajectories when using this model (figure 9.2.3B). For comparison, the covariance of the observed layer trajectories derived based on the GICC05 annual layer boundaries is found in figure 9.2.2B.

When using this model, most of the variability of the individual annual layers from the mean annual layer trajectory is caused by a high value of the white noise component on the layer shape ($\sigma_\varepsilon^2 = 0.35$, $\Phi = 0.05$, $\boldsymbol{\varphi} = 0.65$). This is seen as the line of high variance levels cutting across the middle of the covariance matrix. The variability of the layer trajectories themselves is small. When assuming a cosine as shape function, a high peak will always be connected to a deep trough. Even when allowing the peak heights to vary from layer to layer, this relationship is postulated to hold for all layers, hence giving rise the checkerboard pattern vaguely present in the plot of the covariance pattern in figure 9.2.3B. However, such correspondence between peak and trough amplitudes does not hold for the observed annual layers, whose variance therefore must be described by other means. The variability around the mean annual layer curve must therefore to a large degree be explained as independent noise on the individual observations, i.e. white noise.

The above illustrates that the performance of the layer detection algorithm entirely depends on the selection of an appropriate layer signal model. Having asked it to search for "cosine-like" layers, this is exactly what the algorithm finds. In this case, such assumption is just too much of an approximation.

9.2.2 A more complex cosine-based trajectory

From the previous section it is seen that a plain cosine function is not able to reproduce the layering in the visual stratigraphy data well enough to detect also the more abnormally shaped annual layers. The next question arising is therefore: How can the annual layers be described in a more elaborate fashion, which allows the individual layers to conform to a more variable shape?

For this purpose, the following was selected as layer trajectory function (d is duration of the layer):

$$f(x) = A\cos\frac{2\pi x}{d} + B\left(x - \frac{1}{2}\right) + C, \quad x \in [0, 1]$$

By adding the second term, a straight line crossing zero at $x = 0.5$, the layers are allowed to have different peak height at either side. The advantage of the above formulation is that the average value of the parameter B can be tied to a value of zero: For reasons of symmetry, equally many layers will have a high starting value and a low ending value as the opposite. By furthermore adding a constant term, it has been taken into account that most layers have a relatively sharply defined peaks, and broader valleys. Hence, even if a running mean is subtracted during preprocessing of the data, a fitted cosine curve will generally have a mean value that is slightly positive.

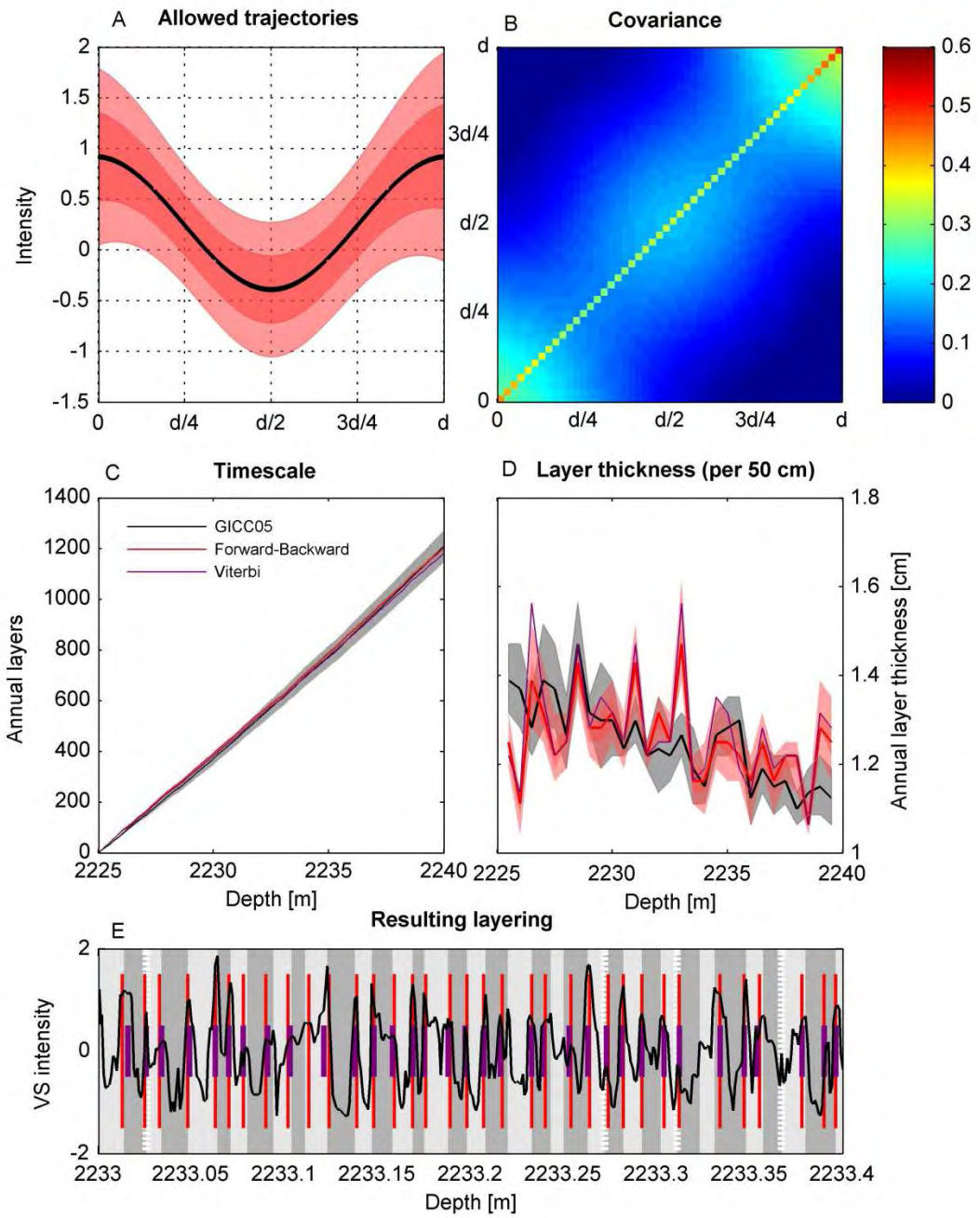


Figure 9.2.4: *Results when using a cosine function combined with a linear function. A, B: Mean trajectory and covariance matrix used as input to the algorithm. C, D: Inferred timescale and derived layer thicknesses. E: An example of the interfered layering in a small section. The background banding shows the GICC05 layers, uncertain layer boundaries are marked with white.*

The increased flexibility of the layer shape trajectories gives rise to a covariance pattern, which, with its broad elongated shape (figure 9.2.4B), is fairly similar to that found for the GICC05 layers. Due to the generally improved fitting of the layer shapes, more of their variability can be attributed a different shape of the layer trajectory. The estimated degree of white noise on the data series is hereby decreased.

As it is seen from figure 9.2.4C, the improvement in performance of the algorithm is evident. In fact, throughout the entire interval, the discrepancy relative to the number of counted layers in GICC05 is merely 5 layers, corresponding to a relative discrepancy of just 0.4%, and the two age estimates are within the uncertainty bands of both. The similarity between the two can also be seen from the evolution of the layer thicknesses with depth. When using this model to describe the annual layers, the modeled annual layers – despite having as input a relatively high annual layer thickness as corresponding to the first half meter of the section – are displaying the same decreasing trend with depth as the GICC05 annual layer thicknesses.

However, not all annual layers are paired up one to one. Some GICC05 layer boundaries are lacking, while other layer boundaries counted in the visual stratigraphy are not indicated as boundaries in GICC05 (an example can be found in the VS data at a depth of 2233.4 m). In total, 9.7% of the layer boundaries cannot be paired up, i.e. averaging to ~5% in each data series (table 9.2.1). But on average there is an equal amount of each, and the resulting timescales match up quite nicely. Also the value of $\Delta/\lambda_{eff} = 0.13$ is small, meaning that the inferred layer boundaries are placed much similar to those in GICC05. It is, however, no surprise that the layer boundaries which are not found, are not taken to be such boundaries. Mainly, this happens at locations where the annual layer signal in the visual stratigraphy is doubtful or non-existing.

9.2.3 A polynomial trajectory

Yet, some things point towards areas of improvement. None of the described cosine based annual layer models were able to reproduce the sharp peaks and broad troughs of the observed annual layers in the visual stratigraphy intensity profile. In order to allow the annual layer model to conform better to such a shape, a polynomial layer trajectory model was tested:

$$f(x) = A\left(x - \frac{1}{2}\right)^2 + B\left(x - \frac{1}{2}\right) + C, \quad x \in [0,1]$$

Again, this particular choice of second order polynomial was chosen for symmetry reasons: It is a good approximation also here to tie the mean value of B to 0, hence allowing equally many high peaks in the beginning as in the end of a layer.

It can be seen from figure 9.2.5A that this type of model is better at recreating the general annual layers observed in the visual stratigraphy data. Its covariance along the trajectory also has an expression which is fairly similar to what has been observed.

Although the obtained timescale does not follow the GICC05 chronology as close as the elaborated cosine function described in the previous section, the two are seen to mainly deviate in the last part of the profile, where the GICC05 layer thicknesses are decreasing faster than what is inferred by the algorithm (figure 9.2.5C,D). Especially in the first 5 meters of the considered section, the GICC05 and the modeled layer chronology follow each other quite nicely, and throughout the section, the inferred number of annual layers is within the maximum counting error band of the manually counted timescale.

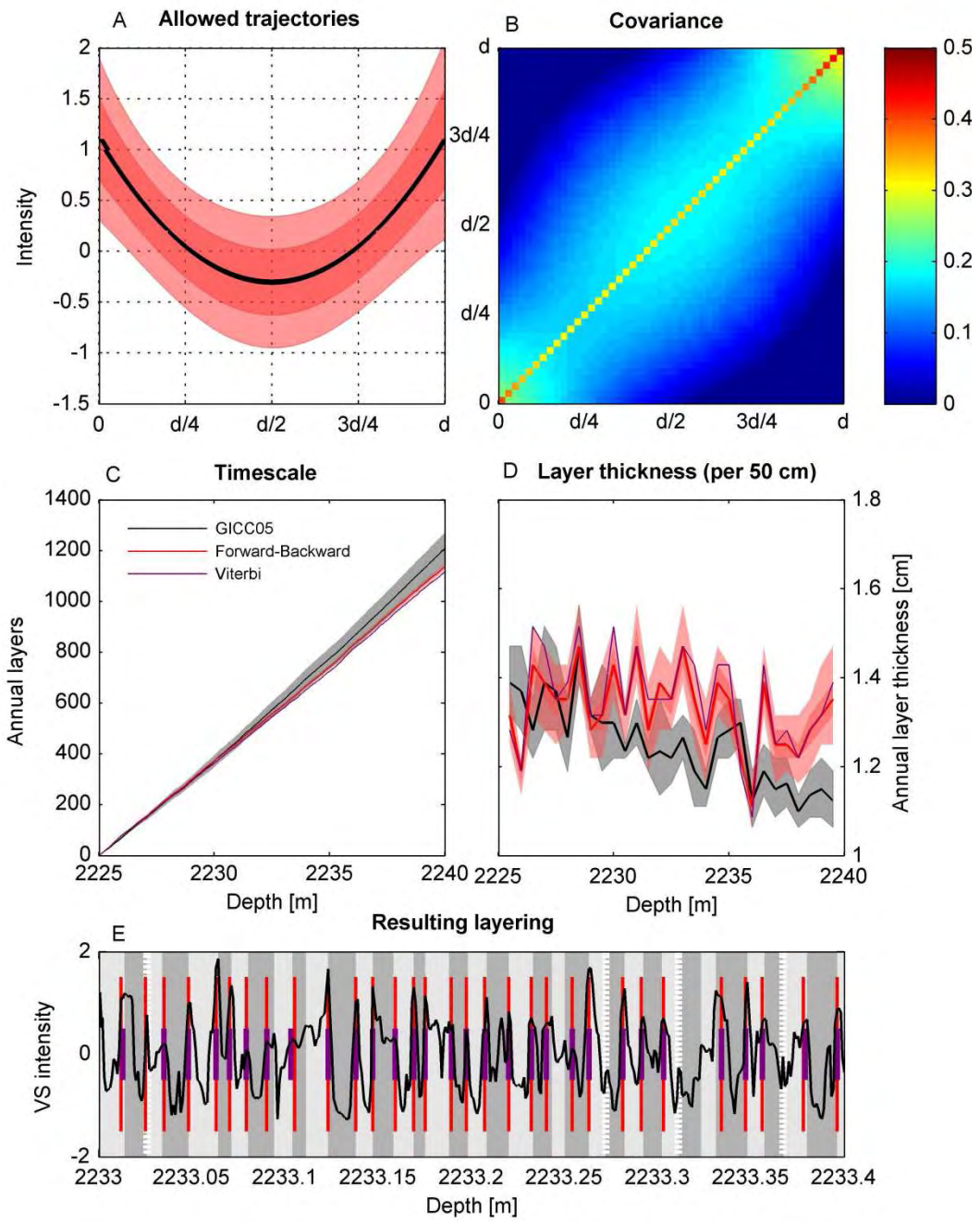


Figure 9.2.5: *Results when using a second order polynomial as trajectory function. A, B: Mean trajectory and covariance matrix used as input to the algorithm. C, D: Inferred timescale and derived layer thicknesses. E: An example of the interfered layering in a small section. The background banding in gray colors shows the GICC05 layers, uncertain layer boundaries are marked with white.*

The much improved shape of the modeled annual layer trajectories with sharp and more well-defined peaks, results in the exact location of the layer boundaries to be determined more precisely with this model than the previous ones, which is seen as a decrease in the value of Δ/λ_{eff} to 0.12.

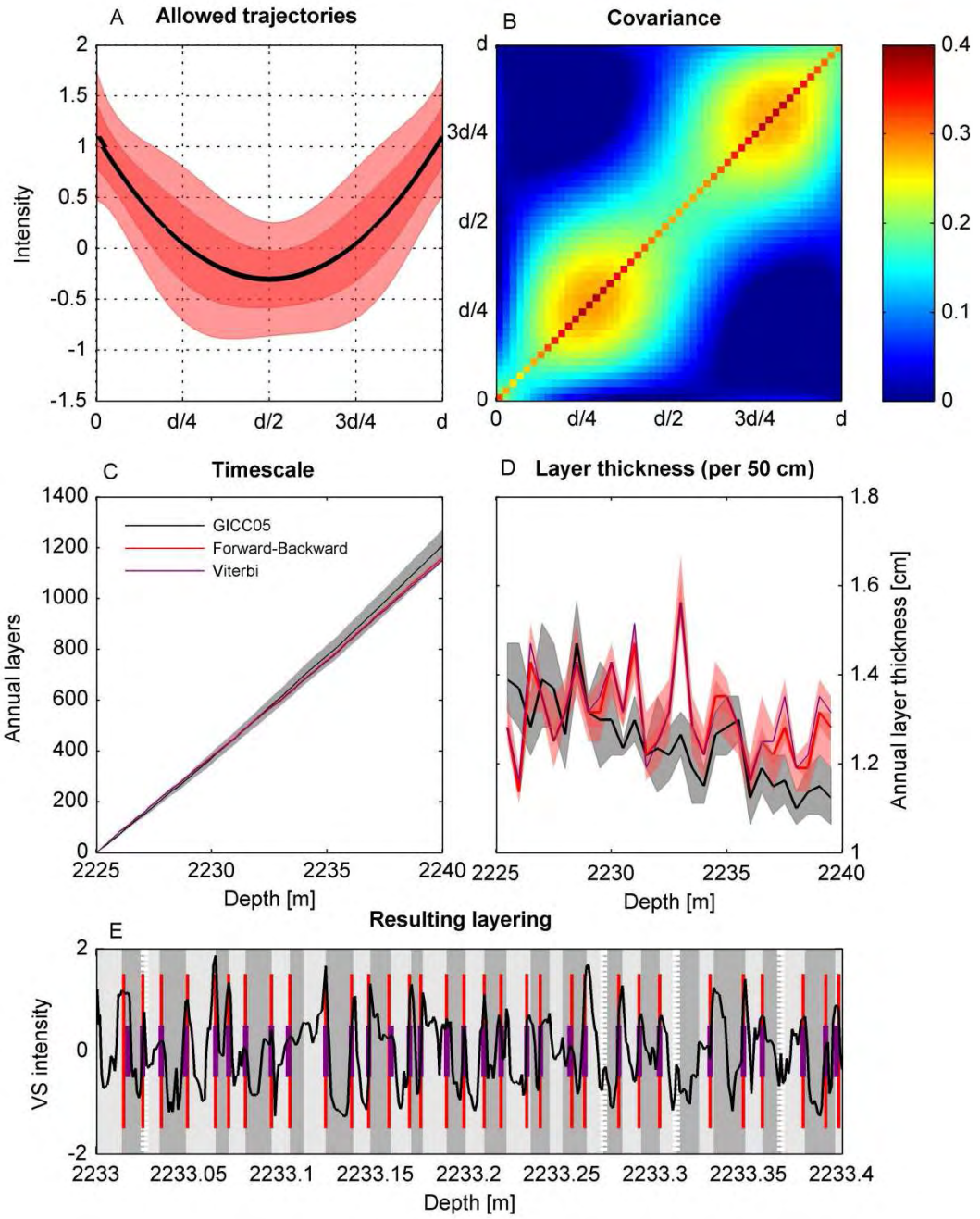


Figure 9.2.6: *Using a third order polynomial as trajectory function. A, B: Mean trajectory and covariance matrix used as input to the algorithm. C, D: Inferred timescale and derived layer thicknesses. E: An example of the interfered layering in a small section. The background banding shows the GICC05 layers, uncertain layer boundaries are marked with white.*

Also a third order polynomial was tested out. The results obtained for this model (figure 9.2.6) were very similar to those derived when using a second order polynomial. Although the model does count a few layers more than found by the second order polynomial model, it still tends to count too few layers compared to GICC05. But again, the discrepancy between the two is mainly within the last couple of meters of the section in consideration.

With increasing complexity of the annual layer model, the model has a better chance to fit the full spectrum of possible layer expressions in the data series. But it happens at the expense of an increasing number of tunable model parameters. In case such parameters are not known beforehand, it will cause the counting results to be less reliable, as they may depend on the exact choice of parameter values. It seems that a third order polynomial model may not be well-enough constrained based on just a small section of the data, and this causes a slightly strange pattern in the covariance matrix to occur. Such covariance matrix is used as input to the model, even if it may not be the best estimate. As furthermore it seems that the performance of the layer detection algorithm does not increase much by adding the extra terms, the results obtained using this model has not been included in the following.

9.2.4 Comparison of model results

A 1 m section of the inferred layering according to the four different annual layer models is shown in figure 9.2.7. By comparing the exact placement of the layer boundaries, it is clear that the main difference between the individual models is generally not the placement of layer boundaries, where these stand out relatively clearly also from a manual counting prespective. To a great extent, disparities among the models are due to differences in how many layers are counted where the intensity profile does not have a clear layer expression. An example can e.g. be found at a depth of 2232.9 m, where the annual layer expression in the data series is vague. It can also be seen, as mentioned previously, that the polynomial layer models in general are able to place the layer boundaries a tad more accurately than those based on a cosine function. A summary of the results from the individual layer models is included in table 9.2.1.

Figure 9.2.8 shows a comparison between the individual model results of the derived mean annual layer thicknesses in 50 cm sections. By considering the mean annual layer thicknesses in small sections, the amount of detected layer boundaries are integrated, and the results are therefore easier to compare. The mean annual layer thicknesses in the same intervals based on the GICC05 chronology are also shown.

There is a striking resemblance between the individual model results. In particular, when not considering the model based on the pure cosine, which previously was shown to be too simplistic for this purpose, and generally not yielding very good results. These four model results are almost independent. Yet, they all are based on the same preprocessing of the visual stratigraphy intensity profile prior to analysis (hereby enhancing some peaks on the expense of others). And also the fixed model parameters going into the annual layer detection algorithm was based on the same 50 cm of layers in the beginning of the entire section. But the model parameters themselves in each of the three models are different, and so is the shape that they are looking for in the data.

Their similarity does pose the question whether a fresh manual layer counting exercise in e.g. the interval around 2233 m would yield a different result with fewer layer boundaries detected. However, it may also be that within this interval, there are simply more layers with a vague or ambiguous expression, which the layer models employed here have not been able to pick up. And one should always remember that the manual counting also

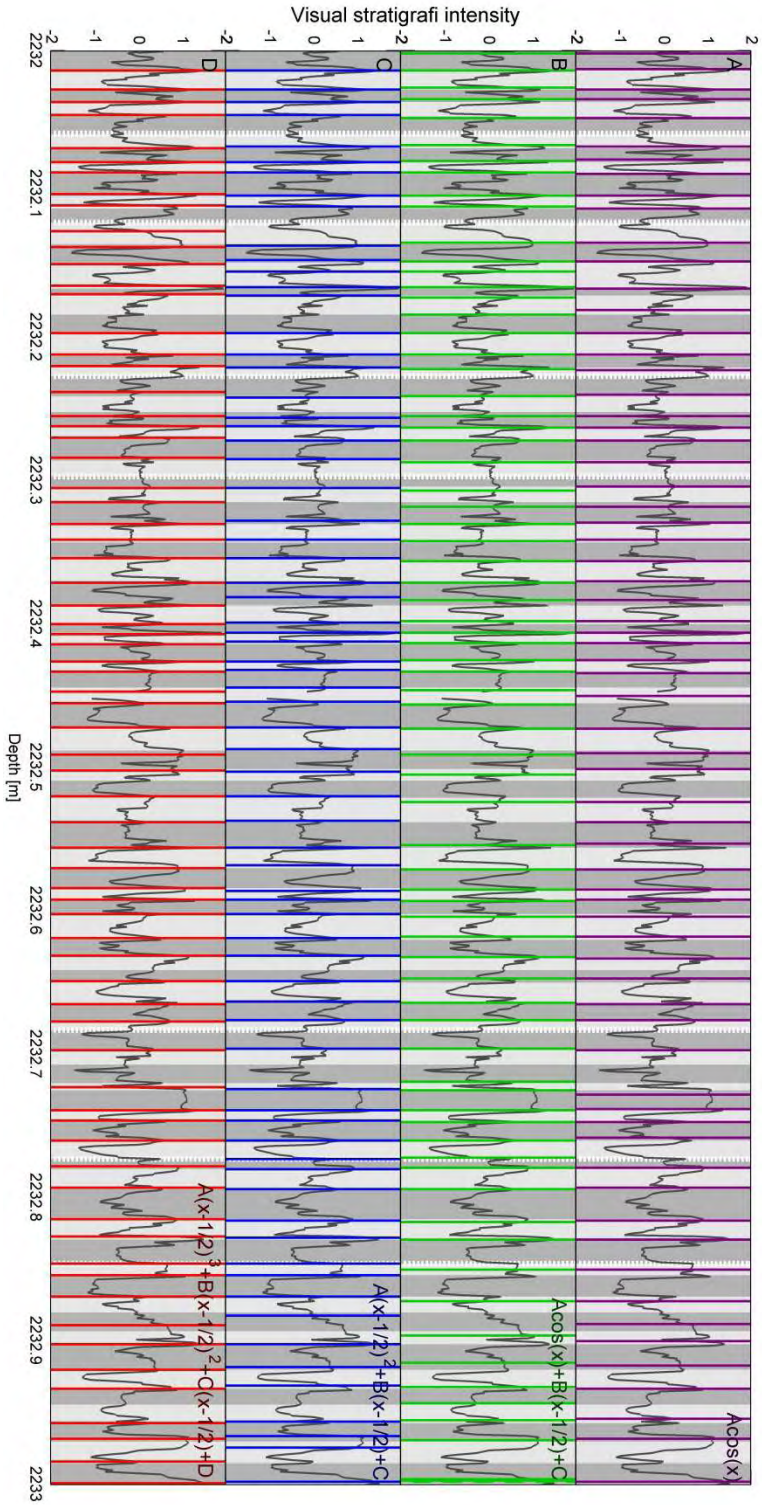


Figure 9.2.7: *The inferred annual layering according to the four different layer models from the depth interval 2232-2233m. A: Simple cosine. B: Cosine plus a first order polynomial. C: Second order polynomial. D: Third order polynomial. The bright and dark gray banding in the background marks the GICC05 layer boundaries, with white stripes being uncertain layer boundaries.*

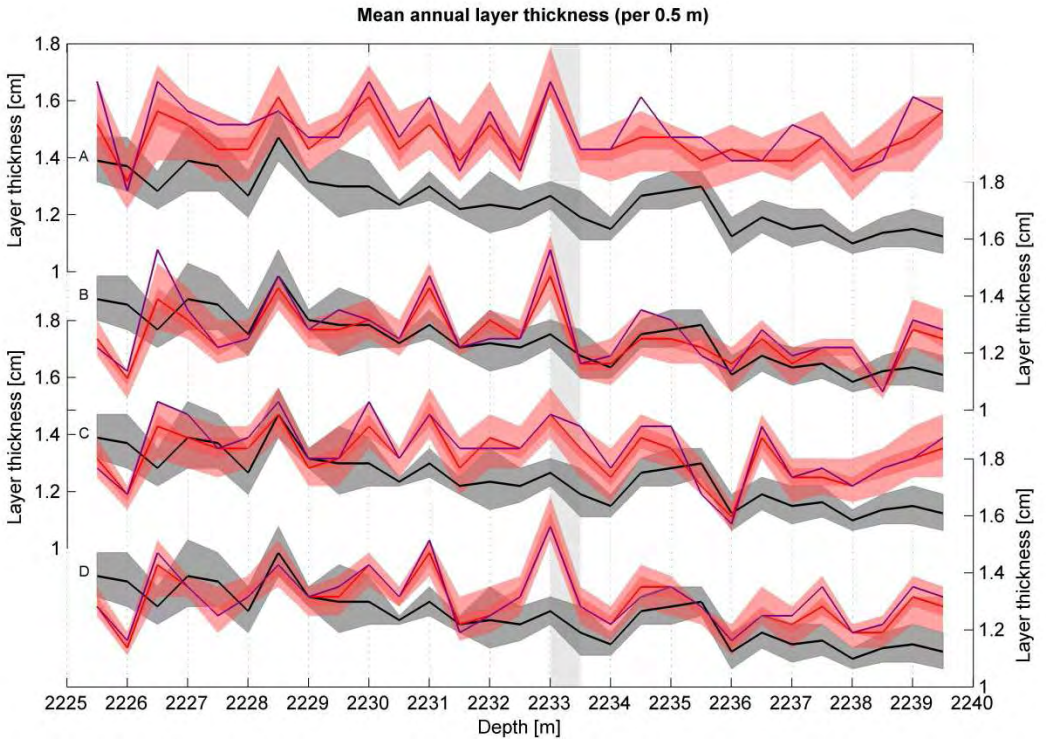


Figure 9.2.8: *Comparison between derived mean annual layer thicknesses in sections of 50 cm for the selected depth interval between 2225-2240 (cold period). A: Simple cosine. B: Cosine plus a first order polynomial. C: Second order polynomial. D: Third order polynomial. For all, the black line is the resulting layer thicknesses based on GIOCC05, and the gray band is the Maximum Counting Error. The vertical band in light gray band show the location of the layering example in figure 9.2.3E, figure 9.2.4E, figure 9.2.5E, and figure 9.2.6E.*

used information from other data records than the visual stratigraphy, whereas the layering inferred here only is based on the visual stratigraphy data.

From figure 9.2.8, the uncertainty with which the annual layer models have estimated the annual layering can be seen. These uncertainties assume the correctness of the annual layer model and should be regarded as a lower limit of the uncertainties. To alleviate this dependency on the model, a more correct estimation of the uncertainties ought to be based on a collection of annual layer models.

The annual layer model based on the cosine has a quite large uncertainty band, implying that the model often cannot conclude with certainty whether something is a layer or not. This can also be seen from the somewhat larger differences between the results obtained from the Forward-Backward algorithm and the Viterbi algorithm. This is due to the cosine being a too tight and simplistic description of how an annual layer is expressed in the data. The remaining models, with their higher degree of flexibility in the allowed annual layer trajectories, estimate the involved uncertainties to be somewhat smaller.

The inferred uncertainty intervals on the annual layering within each little section cannot be fully compared to the Maximum Counting Error (MCE) estimate of the GICC05 chronology. First of all, the MCE is a conservative estimate of the involved uncertainties, and although it can be regarded as a 2σ-error bound (and as such is comparable to the larger of

Trajectory function	$A\cos x$		$A\cos x + B\left(x-\frac{1}{2}\right)+C$		$A\left(x-\frac{1}{2}\right)^2 + B\left(x-\frac{1}{2}\right)+C$	
	Forward-Backward	*Viterbi*	*Forward-Backward*	*Viterbi*	*Forward-Backward*	*Viterbi*
Δ [cm]	0.21	0.19	0.18	0.17	0.17	0.17
Δ/λ_{eff}	0.15	0.13	0.13	0.12	0.12	0.12
F [%]	11.7	12.4	9.7	10.3	10.5	10.9
N	1031	1002	1199	1177	1136	1115
Q_{50}	1026-1035		1195-1203		1132-1140	
Q_{95}	1018-1044		1186-1212		1124-1148	
GICC05	1201±62 years		1204±62 years		1206±62 years	

Table 9.2.1: *Performance of three of the layer models during GS-13: The cosine, the cosine plus first order polynomial, and the second order polynomial. Results using the Forward-Backwards as well as the Viterbi algorithm are noted. The x'te % confidence interval is denoted by Q_x. As the algorithm has not determined the layering in the very last part of the data section, 'the last part' being determined separately for each model, the number of GICC05 layers, which the result should be compared to, varies slightly from model to model. λ_{eff} is calculated based on the GICC05 data for the entire interval, $\lambda_{eff}^{GICC05} = 1.51$ cm. As described in chapter 7, F does not include the fraction of layers being discarded due to uncertain years in the GICC05 chronology.*

the two red bands), it takes into account that the manual counting may be biased, such that the uncertainty estimates of individual layer boundaries does not partly balance out. The layer detection algorithm, on the other hand, assumes the counting to be unbiased. For each section, the individual uncertainties are allowed to partly balance each other out, thereby causing the resulting uncertainty estimate to grow slower with distance.

From the above, it is not clear what should be selected as the best model. Apart from the pure cosine, which bears all the indications of being just too simple, all the employed annual layer models do a decent job, and end up within the Maximum Counting Error estimate of the GICC05 annual layer counting.

The layer trajectory model composed of a cosine plus linear function shows the highest skill, and it estimates almost exactly the same number of annual layers as estimated in the GICC05. However, the model itself is not flawless. The mean annual layer trajectory does not imitate the one found on the basis of the real data very well. A second or third order polynomial is able to reproduce the observed shape much better. Yet, none of the models seem to be fully able to capture the observed covariance between individual data points within a year, and in the future, more time should go into the search for appropriate annual layer models. In spite of these limitations of the individual layer trajectory models investigated here, the obtained results turn out to be quite robust among the respective models.

9.3 Layer detection during a warm period

The depth interval between 2200-2220 m in the NGRIP ice core covers the last part of Greenland Interstadial 12 (GI-12). Within this period, the climate appears to have been relatively stable, and the annual layer thicknesses, as well as the general expression of an annual layer in the line-scan data, are more or less constant. Also here, it is therefore possible to run the layer detection algorithm with all annual layer parameters fixed. These

parameters have been determined based on data from the first part of the interval. However, due to the increased layer thickness in this interval, robust parameter estimates could not be determined based on the first half meter alone. For this reason, layer parameter estimates were instead determined based on the observed layering in the first meter of the data.

The annual layer signal in the visual stratigraphy is much more difficult to recognize during the warm periods than during the cold periods. In figure 9.3.1 is shown the mean layer shapes and corresponding covariance of the annual layers in this interval. Compared to the cold period, the shapes are more leveled with relatively small peaks at the boundaries, and the covariance between individual observations is decreasing faster with distance, and in that way resembling white noise more than what was the case during the cold period. Both of these factors cause annual layer pattern matching to be more difficult within this interval.

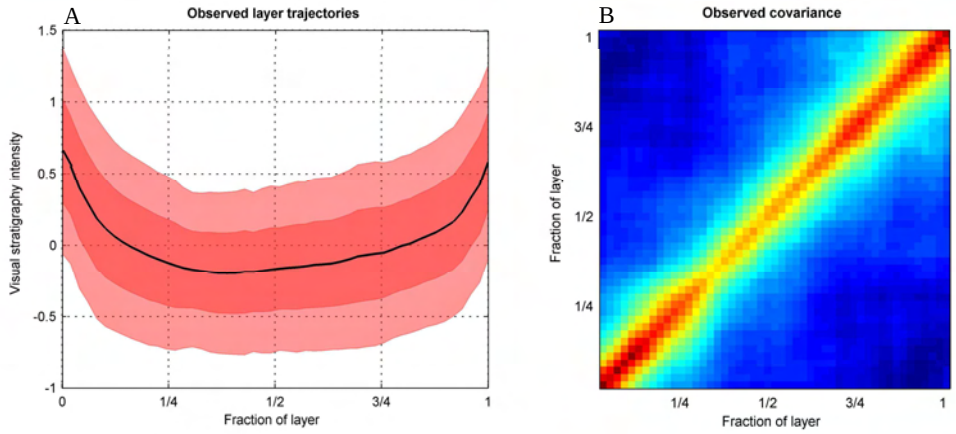


Figure 9.3.1: *A: Mean annual layer trajectories in the visual stratigraphy data in the depth interval from 2200-2220m, covering the last part of GI-12. Red bands are the 1 and 2σ sample covariance bands. B: Covariance corresponding to the annual layer trajectories.*

9.3.1 A simple cosine as trajectory function

The first annual layer model to be considered is again a plain cosine function. The results are shown in figure 9.3.2. Also in this case, the assumption of a cosine is a very rigid conjecture, which gives rise to a of a strange postulated covariance pattern between individual observations within a layer (figure 9.3.2B).

When considering the resulting timescale, it is seen that although the annual layer detection with this layer model as input always is counting too few layers, it always stays within the maximum counting error of the GICC05 chronology. Indeed, it seems that the algorithm is able to point out just about all the certain layers, but that none of the uncertain layers are selected. The end result is that the algorithm is counting about 4.5% layers less than the GICC05, but as mentioned earlier, this is within the estimated maximum counting error band.

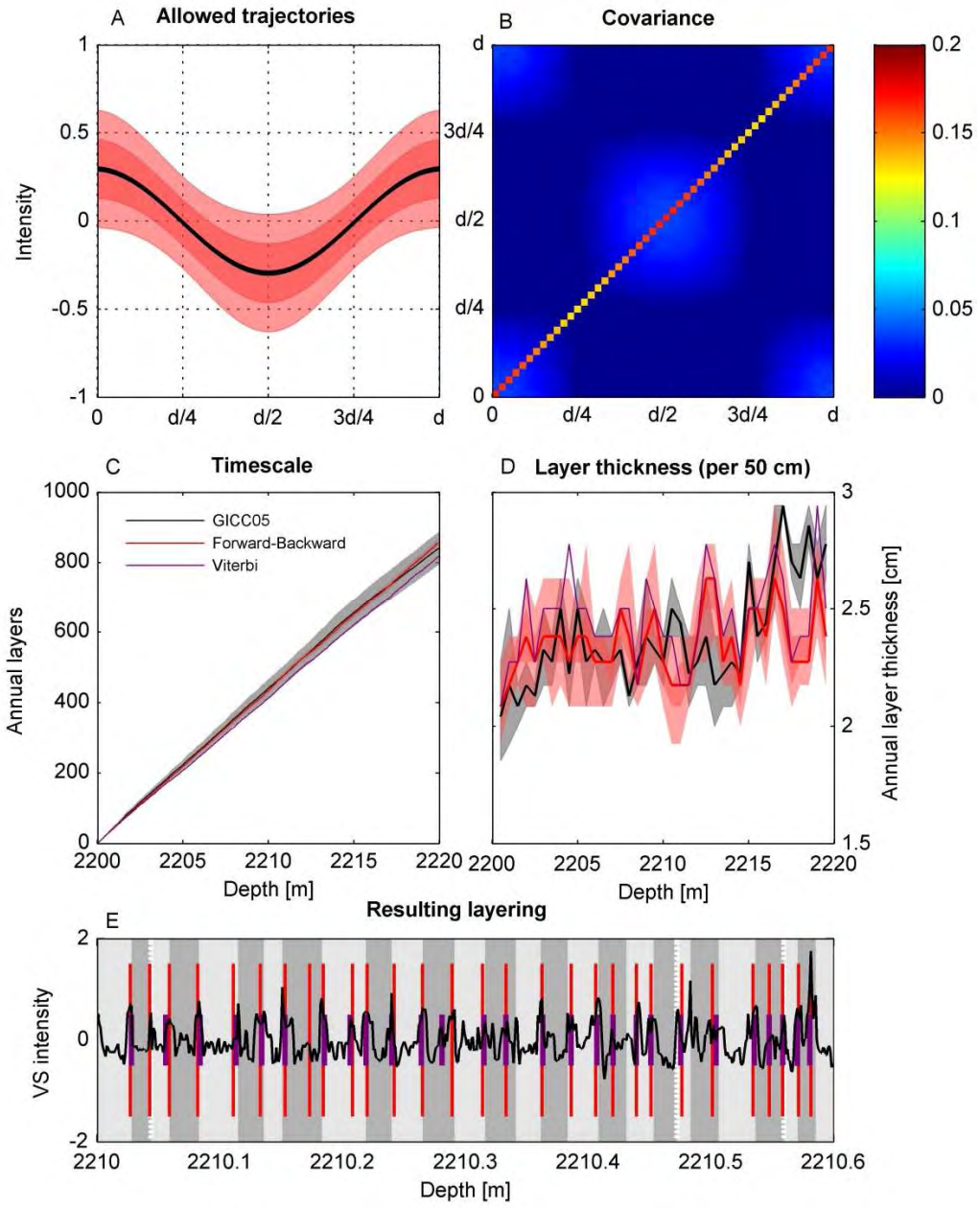


Figure 9.3.2: *Using a cosine as trajectory function. A, B: Mean trajectory and covariance used as input to the algorithm. C, D: Inferred timescale and derived layer thicknesses. E: An example of the interfered layering in a small section. The background banding shows the GICC05 layers, uncertain layer boundaries are marked with white.*

9.3.2 A more complex cosine-based trajectory

Subsequently, a trajectory function consisting of the cosine plus a first order polynomial was considered. The results are shown in figure 9.3.3. This layer model, which showed high skill during the cold periods, seriously overestimates the number of layers within the warm periods. Apparently, the layer shape is too flexible, allowing too many random peaks to be counted as layer boundaries. The total number of inferred annual layers is 25%

above the number estimated by GICC05. This is much beyond the uncertainty on the manual counting, and the inferred results can therefore be rejected.

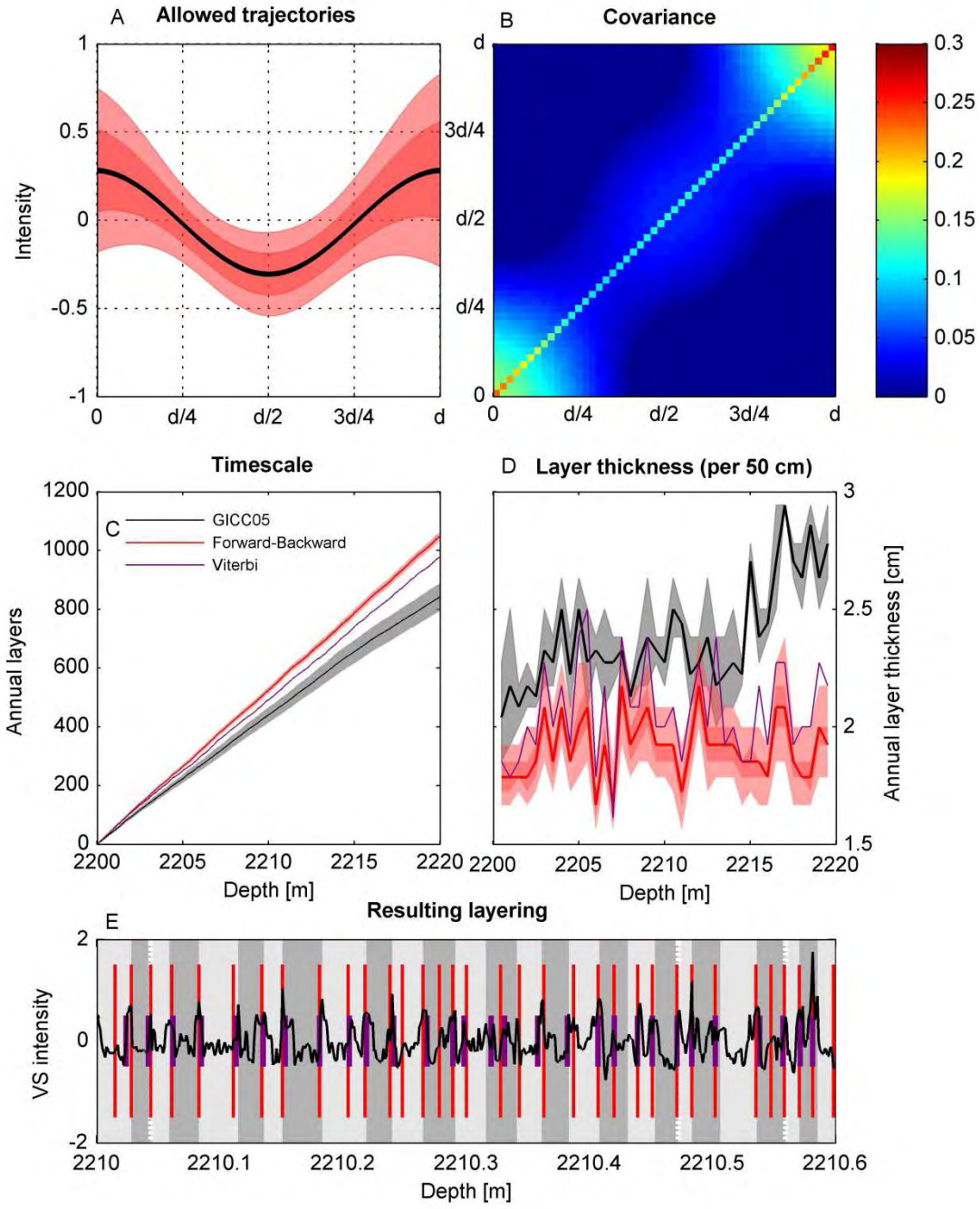


Figure 9.3.3: *Using a cosine plus a linear function as trajectory function. A, B: Mean trajectory and covariance used as input to the algorithm. C, D: Inferred timescale and derived layer thicknesses. E: An example of the interfered layering in a small section. The background banding shows the GICC05 layers, uncertain layer boundaries are marked with white.*

9.3.3 A polynomial trajectory

Just as it was the case for the layer shape based on a cosine plus a linear function, also a second-order polynomial seem to allow too much variety in the individual layer shapes, and hence too many peaks are counted as layers. The number of inferred layers is a little less than for the cosine-based trajectory function, but still leading to an over-estimation of 20%. The same was the case when using a third order polynomial, for which the layer shape is even more flexible.

A next question which emerges may then be whether or not this over-estimation can be due to a wrong choice of value of the model parameters. The model parameters determine how much of the variability within the layer shapes are allowed, and hence have large importance for which peaks should be considered as mere peaks, and which ones should be considered layer boundaries. Could it be that the layers within the first meter of data, which were used to select the employed model parameters, happened to be very variable, and therefore did not constrain the model properly? To investigate the effect of the chosen set of model parameters on the outcome of the layer detection model, the first 5 m of data was used to select a fixed set of model parameters. However, the result turned out almost exactly the same with this new set of model parameters.

9.3.4 Comparison of model results

In table 9.3.1, the obtained results on the inferred layering during GI-12 are summarized for the different trajectory functions. An example of the inferred layering is shown in figure 9.3.5.

The evolution in the obtained mean annual layer thicknesses, based on 50 cm sections of data is shown in figure 9.3.4. The cosine function, which did a decent job doing layer detection for this depth interval, is different from the rest, and much more alike the GICC05. However, the outcomes of the remaining annual layer models are very similar. Due to the general lack of agreement with GICC05 on the total number of annual layers, this similarity should not be interpreted as a miscounted section of the GICC05. Indeed, it implies that all three layer models are counting more or less the same peaks – but not that all the peaks counted are annual layers. Rather, it is a sign that the number of peaks in the visual stratigraphy, which potentially could be annual layers, is high. Hence, layer detection in this interval is very demanding on the appropriateness of the applied annual layer model and/or a better preprocessing, which is able to enhance the ‘correct’ peaks, and suppress peaks which are not related to the seasonal signal.

Trajectory function	$A\cos x$		$A\cos x + B\left(x-\frac{1}{2}\right)+C$		$A\left(x-\frac{1}{2}\right)^2 + B\left(x-\frac{1}{2}\right)+C$	
	Forward-Backward	*Viterbi*	*Forward-Backward*	*Viterbi*	*Forward-Backward*	*Viterbi*
Δ [cm]	0.40	0.4	0.32	0.34	0.29	0.30
Δ/λ_{eff}	0.16	0.16	0.13	0.14	0.12	0.12
F [%]	9.6	10.9	13.4	11.8	11.3	10.6
N	856	817	1046	977	1008	948
Q_{50}	851-860		1041-1052		1003-1013	
Q_{95}	844-868		1032-1061		994-1022	
GICC05	839±47 years		840±47 years		840 ± 47 years	

Table 9.3.1: *Performance of the layer models during GI-12: The cosine, the cosine plus first order polynomial, and the second order polynomial. Results using the Forward-Backwards as well as the Viterbi algorithm are noted. The x'te % confidence interval is denoted by Q_x. λ_{eff} is calculated based on the GICC05 data for the entire interval, $\lambda_{eff}^{GICC05} = 2.48$ cm.*

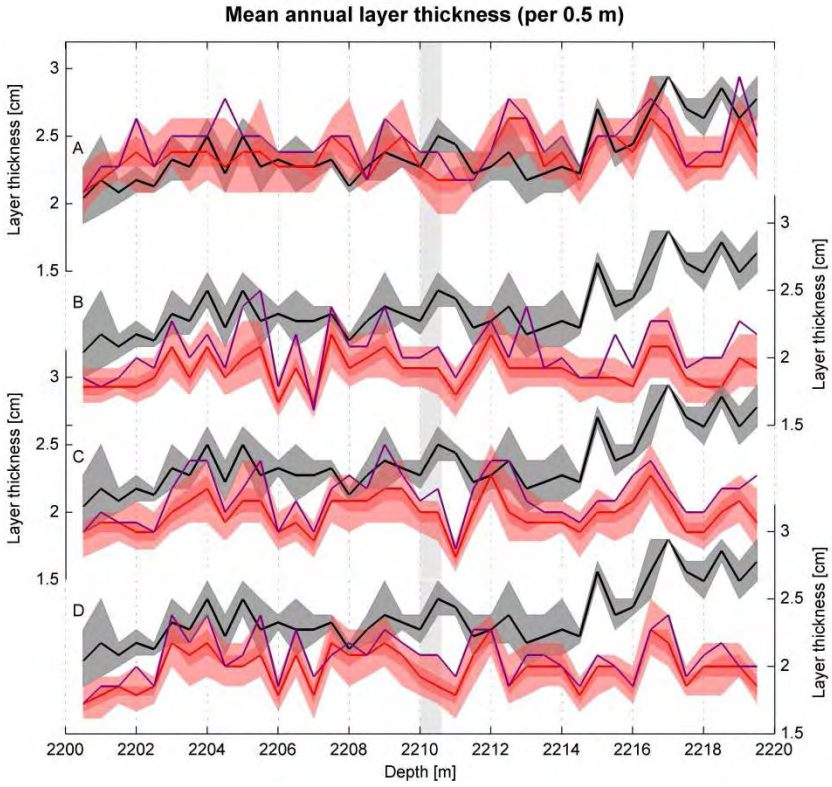


Figure 9.3.4: *Comparison between derived mean annual layer thicknesses in sections of 50 cm for the selected depth interval during GI-12. A: Simple cosine. B: Cosine plus a first order polynomial. C: Second order polynomial. D: Third order polynomial. For all, the black line is the resulting layer thicknesses based on GICC05, and the gray band is the Maximum Counting Error.*

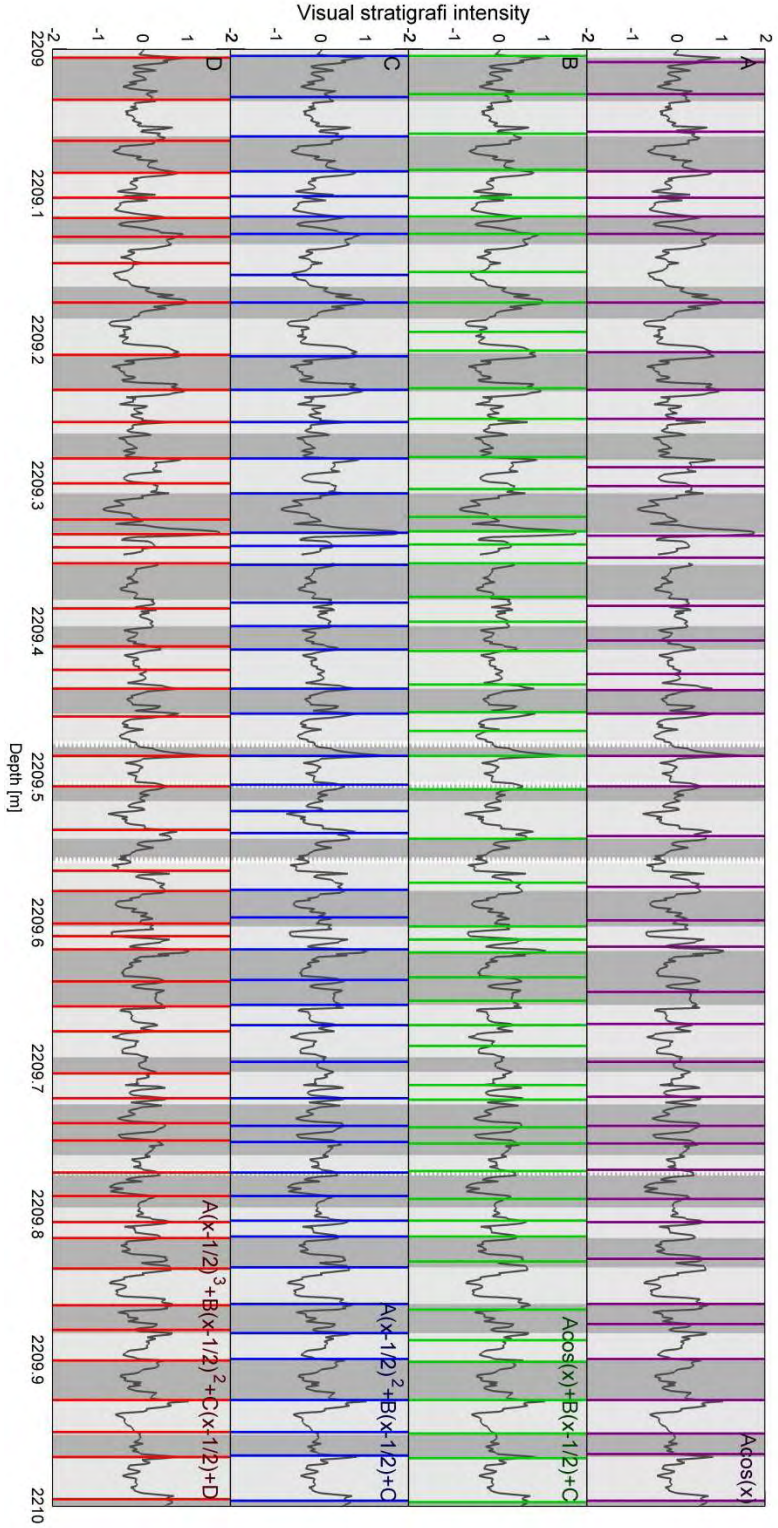


Figure 9.3.5: *A small section of the inferred annual layering during GI-12 for the four investigated annual layer models. A: Simple cosine. B: Cosine plus a first order polynomial. C: Second order polynomial. D: Third order polynomial. The bright and dark gray banding in the background marks the GICC05 layer boundaries, with white stripes being uncertain layer boundaries.*

9.4 Layer detection during onset of GI-12

Finally, the layer detection algorithm was tested over a transitional period from cold to warm, namely over the onset of GI-12. As for the previous tests, the algorithm was run downwards the core, starting in ice deposited during the warm interstadial (with large annual layer thicknesses), and towards the older and deeper ice deposited during the stadial (small layer thicknesses). From GICC05, the layer thicknesses across the transition is known to change with more than a factor two, and also the visual stratigraphy data changes dramatically over the transition. Hence, it is not an adequate approximation to consider the parameters describing the layer thickness distribution or the mean layer signal in the data series as constant across the transition.

Yet again, the algorithm was tested in the simplest mode possible: Only the mean layer thickness was allowed to change with depth. To account for the large evolution also in peak height in the data series across the transition, an extra step was taken in the preprocessing of the visual stratigraphy data before analysis: The data was normalized over 50 cm intervals according to their minimum and maximum values. Apart from a scaling factor, such preprocessing does not change the data at any location much, but it ensures the signal to maintain approximately the same peak heights down the ice core. The visual stratigraphy data changes more with climate, and hence with depth, than what can be rectified by merely adjusting the peak height of the signal. These changes were not taken into account here.

The algorithm was given the best starting point possible. With the algorithm starting out in a warm period, during which the annual layering was best described with a pure cosine function (see section 9.3.1), this was the layer model employed. The parameter input was determined based on the first 2 meters of visual stratigraphy data, during which the annual layer thicknesses were fairly constant.

For each batch, the parameter describing the location parameter of the annual layer thickness distribution (μ_a) was allowed to change. The algorithm was run in Maximum-Likelihood mode, such that no input regarding prior knowledge on the layer thickness distribution was used. The algorithm was allowed to iterate 10 times for each batch, at which point it was assumed that it had reached convergence, and a next step was taken. The result is shown in figure 9.4.1.

The result is remarkably good: The layer detection algorithm manages to adapt to the changing environment and find an appropriate value of the mean annual layer thickness throughout the transitional zone. The inferred timescale can be found in a larger format in figure 9.4.2. In fact, the algorithm has most problems in the beginning and end of the interval, where the algorithm does not locate enough layers. However, this under-counting should come as no surprise. It was exactly the same as what was found when running this extremely simplified model separately for the cold and warm periods before and after respectively.

In total, the 95% confidence interval for the number of annual layers within the transitional period investigated is [257, 273], which should be compared to the GICC05 estimate of

278 $\pm$ 12 layers within the same interval. The two counting intervals are nicely overlapping.

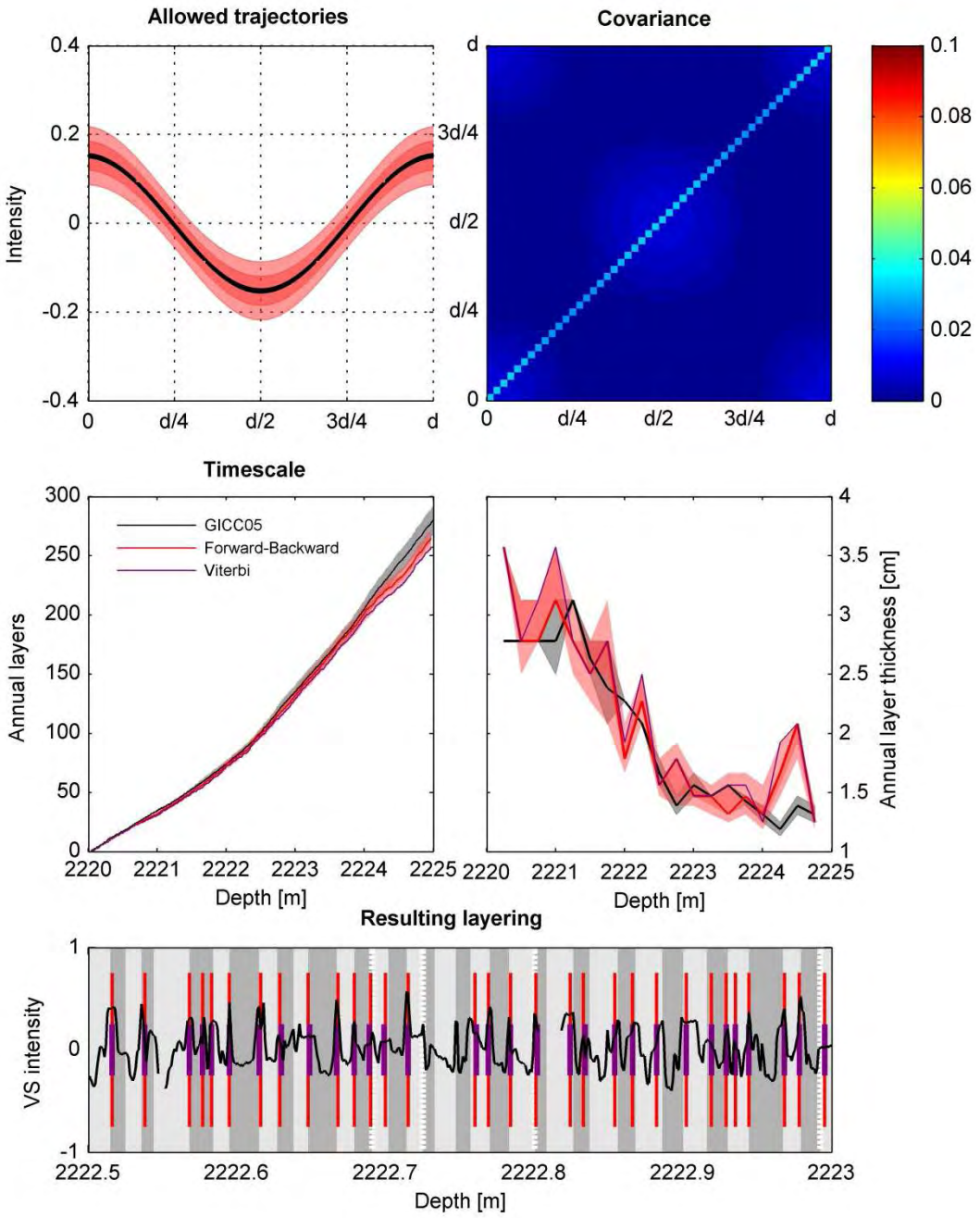


Figure 9.4.1: *Using a cosine as trajectory function over the onset of GI-12. A, B: Mean trajectory and covariance used as input to the algorithm. C, D: Inferred timescale and derived layer thicknesses. E: An example of the interfered layering in a small section. The background banding shows the GICC05 layers, note the uncertain layer boundaries which are marked with white.*

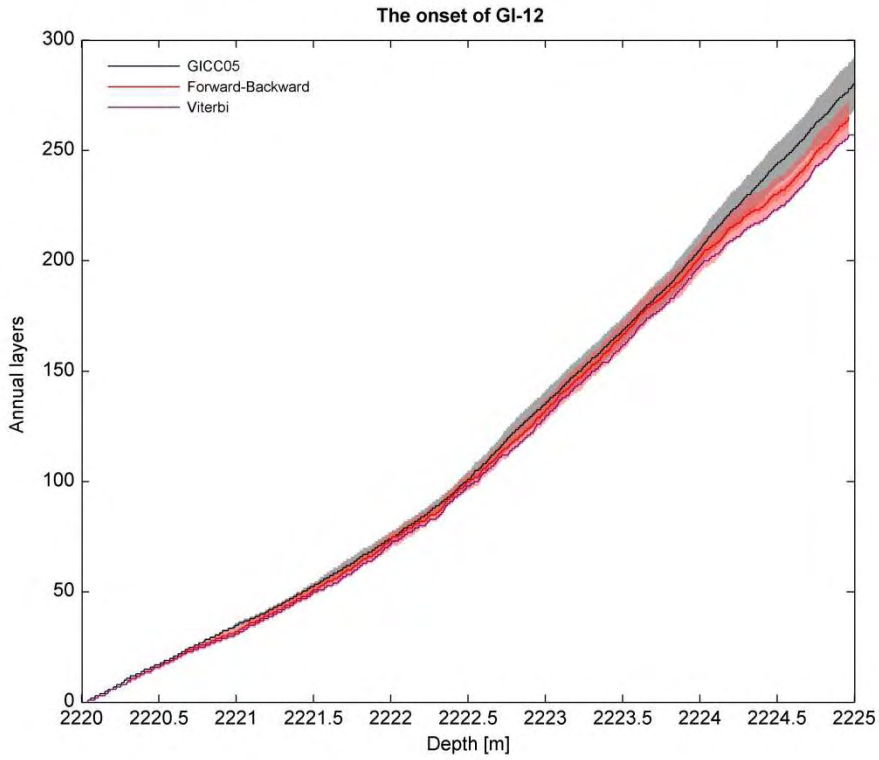


Figure 9.4.2: *The inferred timescale and accompanying uncertainties over the onset of GI-12 (red), and the manually counted GICC05 timescale with the Maximum Counting Error uncertainty band (black/gray).*

9.5 Next steps for development of algorithm

The annual layer detection algorithm based on Hidden Markov Modeling is not yet fully developed. To take it to the next level, where it is able to produce a timescale on its own, further investigations are needed. The major issues to be considered include: How should the data series be processed before analysis to enhance the annual layer signal? What kind of annual layer model should be chosen? And when having selected a model: How do the parameter values vary with depth and with climatic conditions? Longer data sections from many different time periods should be investigated in order to find an answer to these questions.

Although the algorithm has been implemented to be run in Maximum a Posteriori mode, this has not been used in the above. In its present form, this part of the algorithm seems not entirely stable. Most likely, this is due to poor knowledge on the parameter values $\boldsymbol{\varphi}$, towards which the sensitivity studies showed that the algorithm was particularly sensitive. Although Maximum a Posteriori theoretically is the most beneficial way to run the algorithm, it has one important drawback: By allowing prior information on the parameter values to be taken into account, it requires that we have sufficient knowledge of what should go into such priors. Hence, for a proper incorporation of priors in the algorithm to

make sense, the appropriate values for these priors should be investigated. For these reasons, only Maximum Likelihood estimates were used in the above. That the algorithm even in Maximum Likelihood mode was able to correctly follow the two-fold increase in layer thicknesses, which took place during the onset of GI-12, just proves the strength of the method.

The choice of annual layer model is overwhelmingly important for the algorithm to be able to distinguish between seasonal and random peaks in the data series, and hence for the correctness of the inferred annual layer count. But, as it turned out, it is not easy to choose a layer model, which is flexible enough to locate abnormal layers, but not too flexible such that it counts all peaks. Furthermore; the model should work both during the cold periods as well as during the warm periods. A 'correct layer model' may well be impossible to find, and fortunately, it is not required either. The implemented layer trajectory models are crude and simplistic, and in many ways not reflecting the annuals as observed in the data. In spite of this, they perform quite well – in particular so during the cold periods.

Instead of searching for an optimal annual layer model, a different approach may be taken: The algorithm allows for several criteria to be made concerning an annual layer. This implies that it is possible to use more than just a single layer model. Each layer could be matched to two separate annual layer trajectory functions, and evaluated based on its resemblance to both of these. Yet, by doing so, the dependency of the result on the applied annual layer model can no longer be assessed by comparing the results based on the respective annual layer models.

Nevertheless, regardless of the vast amount of things which can be improved upon when it comes to the specific annual layer model and parameter values that should be employed, this does not change one specific, fundamental issue with the layer detection algorithm as it presently stands: It is still just a single-parameter method. Although the algorithm has been made 'semi-multi-parameter-like' by adding also the derivatives of the observation sequence as extra sequences, they essentially contain same information.

Single-parameter methods have generally proven notoriously difficult to make work properly [*Meese et al.,* 1997]. This is also the case when using the layer detection algorithm on the visual stratigraphy data only. The algorithm does exactly what it is asked to do: Finds peaks in the data series which are the most likely to be peaks connected to the annual cycle. But in cases where even an experienced investigator would be in doubt, because there essentially is not enough information, then so is the algorithm. Another next step is therefore to incorporate the use of several data series. Only a layer detection algorithm, which is able to take into account several data series, will be able to obtain the same (or higher?) accuracy as can be made by comparing and combining data sequences by eye.

In terms of programming, the addition of extra data series is straight-forward. The challenge with the addition of extra data sequences is based on the data itself: First of all, the data series must be co-registered to be on the exact same depth scale. Also, issues regarding data series where diffusion has taken place must be considered. Almost no diffusion has occurred to the visual stratigraphy data, whereas e.g. the conductivity is rather heavily diffused. With diffusion, the assumption of individual layers to be independent, and the

individual observations within a layer to be conditionally independent, is no longer valid (if it ever was). The inclusion of such data series may therefore require the data to be back-diffused before analysis. To some extent, this may allow for the re-establishment of lost features in the data, and thus decrease the dependency between observations belonging to individual layers.

10. Concluding remarks

An accurate chronology is the fundament for a correct interpretation of a paleoclimatic record. In this respect, the Greenland ice cores are unique, in that they allow for very accurate chronologies to be established far back in time. With their high temporal resolution, it has proven possible to establish an annual layer counted chronology reaching back to 60 ka BP. As the subjectivity involved in manual layer interpretation is increasing with depth, the chronology cannot manually be extended further back in time.

Automated procedures for annual layer counting have generally proven notoriously difficult to develop, and with a performance much inferior to manual layer counting. Yet, I believe that the algorithm developed here can represent a first step towards a high-quality automated method of annual layer counting in ice cores. Based on the statistical framework of Hidden Markov Modeling, originally developed for machine speech recognition, it presents a mathematically rigorous yet efficient method to determine the most likely layering in a data series. Its fundamental force lies in the way that the algorithm is able to imitate the manual procedures, while being based on purely objective criteria for annual layer recognition.

In its present form, the annual layer detection algorithm does suffer from a few 'teething troubles'. Due to lack of time to fully investigate for an appropriate description of the appearance of an annual layer in the data, the methodology has not yet been implemented to provide an accurate chronology. However, even with an initial and relatively random guess of layer model, the algorithm proved able to correctly identify the annual layering over the onset of a Dansgaard-Oeschger event with a corresponding halving in annual layer thicknesses over less than five meters.

The layer detection algorithm has here been applied to visual stratigraphy data from the NGRIP ice core, in which the annual signal seems to be maintained to great depths. This data series may therefore potentially be used for extending the GICC05 chronology further back in time. However, when using an automated approach it is extremely important that data is not corrupted. The NGRIP visual stratigraphy profile had to go through extensive treatment in order for the layer detection algorithm not to be confused by what it regarded as an annual layer signal randomly changing with depth. In that respect, manual layer counting is much more robust. But even with the above mentioned reconstruction of the

visual stratigraphy profile, the expression of an annual layer in these data is very changeable, and the record as such not very reliable for annual layer detection.

The layer detection algorithm has been developed with the visual stratigraphy data in mind. But it has been developed in a general setting, which allows it to relatively easily be adapted to use for other kinds of annually laminated data. One of the most interesting prospects may be the possible development of the algorithm into a multi-parameter method. Only by taking a multi-parameter approach will an automated method be able to fully compete with manual counting.

Speech recognition software based on HMMs has been developed over more than 40 years. With annual layer detection, we have just started.

Bibliography

Alley, R. B., et al. (1997), Visual-stratigraphic dating of the GISP2 ice core: Basis, reproducibility, and application, *J Geophys Res-Oceans*, *102*(C12), 26367-26381.

Andersen, K. K., P. D. Ditlevsen, S. O. Rasmussen, H. B. Clausen, B. M. Vinther, S. J. Johnsen, and J. P. Steffensen (2006a), Retrieving a common accumulation record from Greenland ice cores for the past 1800 years (vol 111, art no D15106, 2006), *J Geophys Res-Atmos*, *111*(D18), -.

Andersen, K. K., et al. (2006b), The Greenland Ice Core Chronology 2005, 15-42 ka. Part 1: constructing the time scale, *Quaternary Sci Rev*, *25*(23-24), 3246-3257.

Antti, K. (1996), Modelling ECG signals with hidden Markov models, *Artificial Intelligence in Medicine*, *8*(5), 453-471.

Baum, L. E., and T. Petrie (1966), Statistical Inference for Probabilistic Functions of Finite State Markov Chains, *Annals of Mathematical Statistics*, *37*(6), 1554-&.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970), A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, *The Annals of Mathematical Statistics*, *41*(1), 164-171.

Benson, C. S. (1962), Stratigraphic studies in the snow and firn of the Greenland ice sheet *Rep. AD-288 219*.

Bigler, M., A. Svensson, E. Kettner, P. Vallelonga, M. E. Nielsen, and J. P. Steffensen (2011), Optimization of High-Resolution Continuous Flow Analysis for Transient Climate Signals in Ice Cores, *Environ Sci Technol*, *45*(10), 4483-4489.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, xx, 738 p. pp., Springer, New York.

Bradley, R. S. (1985), *Quaternary paleoclimatology : methods of paleoclimatic reconstruction*, xvii, 472 p. pp., Allen & Unwin, Boston.

Bulla, J., and I. Bulla (2006), Stylized facts of financial time series and hidden semi-Markov models, *Computational Statistics & Data Analysis*, *51*(4), 2192-2209.

Chien, J.-T. (2002), Quasi-Bayes Linear Regression for Sequential Learning of Hidden Markov Models, *Ieee T Acoust Speech*, *10*(5), 268-278.

Chien, J.-T., and C.-H. Huang (2003), Bayesian learning of speech duration models, *Ieee Transactions on Speech and Audio Processing*, *11*(6), 558-567.

Dahl-Jensen, et al. (1997), *A search in north Greenland for a new ice-core drill site*, International Glaciological Society, Cambridge, ROYAUME-UNI.

Dahl-Jensen, D., S. J. Johnsen, C. Hammer, H. B. Clausen, and J. Jouzel (Eds.) (1993), *Past accumulation rates derived from observed annual layers in the GRIP ice core from Summit, Central Greenland*, Springer, New York.

Dahl-Jensen, D., N. S. Gundestrup, H. Miller, O. Watanabe, S. J. Johnsen, J. Steffensen, P. rgen, H. B. Clausen, A. Svensson, and L. B. Larsen (2002), The NorthGRIP deep drilling programme, *Annals of Glaciology*, *35*(1), 1-4.

Dansgaard, W., and S. J. Johnsen (1969), A Flow Model And A Timescale For The Ice Core From Camp Century, Greenland *J Glaciol*, *8*(53).

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum Likelihood from Incomplete Data Via Em Algorithm, *Journal of the Royal Statistical Society Series B-Methodological*, *39*(1), 1-38.

Faisan, S., L. Thoraval, J. P. Armspach, J. R. Foucher, M. N. Metz-Lutz, and F. Heitz (2005), Hidden Markov event sequence models: Toward unsupervised functional MRI brain mapping, *Academic Radiology*, *12*(1), 25-36.

Faria, S. H., J. Freitag, and S. Kipfstuhl (2010), Polar ice structure and the integrity of ice-core paleoclimate records, *Quaternary Sci Rev*, *29*(1-2), 338-351.

Fisher, D. A., N. Reeh, and H. B. Clausen (1985), Stratigraphic noise in time series derived from ice cores, *Annals of Glaciology*, *7*, 76-83.

Gauvain, J. L., and C. H. Lee (1994), Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, *Ieee Transactions on Speech and Audio Processing*, *2*(2), 291-298.

Gauvain, J. L., L. F. Lamel, G. Adda, and M. Addadecker (1994), Speaker-Independent Continuous Speech Dictation, *Speech Communication*, *15*(1-2), 21-37.

Gish, H. (1993), A segmental speech model with applications to word spotting.

Gish, H., and K. Ng (1996), Parametric trajectory models for speech recognition, *Fourth International Conference on Spoken Language,*, *1*(ICSLP 96 Proceedings).

Gow, A. J. (1968), Deep core studies of the accumulation and densification of snow at Byrd Station and Little America V, Antarctica*Rep. AD-669 240*, 45 pp.

Gupta, M. R., and Y. Chen (2011), Theory and Use of the EM Algorithm, *Found. Trends Signal Process.*, *4*(3), 223-296.

Hamilton, W. L., and C. C. Langway (1968), A Correlation of Microparticle Concentrations with Oxygen Isotope Ratios in 700 Year Old Greenland Ice, *Earth Planet Sc Lett*, *3*(4), 363-&.

Hammer, C. U. (1980), Acidity of Polar Ice Cores in Relation to Absolute Dating, Past Volcanism, and Radio-Echoes, *J Glaciol*, *25*(93), 359-372.

Hammer, C. U., H. B. Clausen, W. Dansgaard, N. S. Gundestrup, S. Johnsen, and N. Reeh (1978), Dating of Greenland ice cores by flow models, isotopes, volcanic debris, and continental dust, *J Glaciol*, *20*, 3-26.

Huo, Q., and C. H. Lee (1997), On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate, *Ieee T Acoust Speech*, *5*(2), 161-172.

Jelinek, F., L. R. Bahl, and R. L. Mercer (1975), Design of a Linguistic Statistical Decoder for Recognition of Continuous Speech, *Ieee T Inform Theory*, *It21*(3), 250-256.

Johnsen, S. (1977), Stable isotope homogenization of polar firn and ice, paper presented at Isotopes and Impurities in Snow and Firn, Grenoble.

Johnsen, S. J., D. Dahl-Jensen, N. Gundestrup, J. P. Steffensen, H. B. Clausen, H. Miller, V. Masson-Delmotte, A. E. Sveinbjornsdottir, and J. White (2001), Oxygen isotope and palaeotemperature records from six Greenland ice-core stations: Camp Century, Dye-3, GRIP, GISP2, Renland and NorthGRIP, *J Quaternary Sci*, *16*(4), 299-307.

Johnsen, S. J., H. B. Clausen, W. Dansgaard, K. Fuhrer, N. Gundestrup, C. U. Hammer, P. Iversen, J. Jouzel, B. Stauffer, and J. P. Steffensen (1992), Irregular Glacial Interstadials Recorded in a New Greenland Ice Core, *Nature*, *359*(6393), 311-313.

Katsuta, N., M. Takano, T. Okaniwa, and M. Kumazawa (2003), Image processing to extract sequential profiles with high spatial resolution from the 2D map of deformed laminated patterns, *Comput Geosci-Uk*, *29*(6), 725-740.

Kim, S., and P. Smyth (2006), Segmental Hidden Markov Models with Random Effects for Waveform Modeling, *J. Mach. Learn. Res.*, *7*, 945-969.

Kim, S., P. Smyth, and S. Luther (2004), Modeling waveform shapes with random effects segmental hidden Markov models, in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, edited, pp. 309-316, AUAI Press, Banff, Canada.

Kipfstuhl, S., F. Pauer, W. F. Kuhs, and H. Shoji (2001), Air bubbles and clathrate hydrates in the transition zone of the NGRIP deep ice core, *Geophys Res Lett*, *28*(4), 591-594.

Lagendijk, R. L., J. Biemond, and D. E. Boekee (1990), Identification and Restoration of Noisy Blurred Images Using the Expectation-Maximization Algorithm, *Ieee T Acoust Speech*, *38*(7), 1180-1191.

Langway, C. C., Jr. (1967), Stratigraphic analysis of a deep ice core from Greenland*Rep. AD-655 164*, 130 pp.

Lemieux-Dudon, B., E. Blayo, J. R. Petit, C. Waelbroeck, A. Svensson, C. Ritz, J. M. Barnola, B. M. Narcisi, and F. Parrenin (2010), Consistent dating for Antarctic and Greenland ice cores, *Quaternary Sci Rev*, *29*(1-2), 8-20.

Madsen, H. (2008), *Time series analysis*, 380 p. pp., Chapman & Hall/CRC, Boca Raton.

McGwire, K. C., J. R. McConnell, R. B. Alley, J. R. Banta, G. M. Hargreaves, and K. C. Taylor (2008a), Dating annual layers of a shallow Antarctic ice core with an optical scanner, *J Glaciol*, *54*(188), 831-838.

McGwire, K. C., G. M. Hargreaves, R. B. Alley, T. J. Popp, D. B. Reusch, M. K. Spencer, and K. C. Taylor (2008b), An integrated system for optical imaging of ice cores, *Cold Regions Science and Technology*, *53*(2), 216-228.

Meese, D. A., A. J. Gow, R. B. Alley, G. A. Zielinski, P. M. Grootes, M. Ram, K. C. Taylor, P. A. Mayewski, and J. F. Bolzan (1997), The Greenland Ice Sheet Project 2 depth-age scale: Methods and results, *J Geophys Res-Oceans*, *102*(C12), 26411-26423.

Mosegaard, K., and A. Tarantola (1995), Monte-Carlo Sampling of Solutions to Inverse Problems, *J Geophys Res-Sol Ea*, *100*(B7), 12431-12447.

Nielsen, S. W. (2005), Registrering af visuel stratigrafi i NGRIP iskernen: Konstruktion, dataopsamling og analyse., Master Thesis thesis, University of Copenhagen, Copenhagen.

North Greenland Ice Core Project Members (2004), High-resolution record of Northern Hemisphere climate extending into the last interglacial period, *Nature*, *431*(7005), 147-151.

Norton, M. P., and D. G. Karczub (2003), *Fundamentals of noise and vibration analysis for engineers*, 2nd ed., 631 pp., Cambridge University Press, Cambridge, New York.

Ostendorf, M., V. V. Digalakis, and O. A. Kimball (1996), From HMM's to segment models: A unified view of stochastic modeling for speech recognition, *Ieee Transactions on Speech and Audio Processing*, *4*(5), 360-378.

Parrenin, F., J. Jouzel, C. Waelbroeck, C. Ritz, and J. M. Barnola (2001), Dating the Vostok ice core by an inverse method, *J Geophys Res-Atmos*, *106*(D23), 31837-31851.

Parrenin, F., F. Remy, C. Ritz, M. J. Siegert, and J. Jouzel (2004), New modeling of the Vostok ice flow line and implication for the glaciological chronology of the Vostok ice core, *J Geophys Res-Atmos*, *109*(D20).

Parrenin, F., et al. (2007), The EDC3 chronology for the EPICA dome C ice core, *Clim Past*, *3*(3), 485-497.

Pauer, F., J. Kipfstuhl, and W. F. Kuhs (1996), Raman spectroscopic study on the spatial distribution of nitrogen and oxygen in natural ice clathrates and their decomposition to air bubbles, *Geophys Res Lett*, *23*(2), 177-180.

Petersen, K. B., and M. S. Pedersen (2008), The Matrix Cookbook, edited, Technical University of Denmark.

Press, W. H. (1996), *Numerical recipes in fortran 77 : the art of scientific computing*, 2nd ed., xxxi, 933 p. pp., Cambridge University Press, Cambridge England ; New York.

Rabiner, L. R. (1989), A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition, *Proceedings of the Ieee*, *77*(2), 257-286.

Raiffa, H., and R. Schlaifer (1961), *Applied statistical decision theory*, 356 p. pp., Division of Research, Graduate School of Business Adminitration, Harvard University, Boston,.

Ram, M., and M. Illing (1994), Polar Ice Stratigraphy from Laser-Light Scattering - Scattering from Meltwater, *J Glaciol*, *40*(136), 504-508.

Ram, M., M. Illing, P. Weber, G. Koenig, and M. R. Kaplan (1995), Polar ice stratigraphy from laser-light scattering: Scattering from ice, *Geophys Res Lett*, *22*(24), 3525-3527.

Rasmussen, S. O., K. K. Andersen, M. L. Siggaard-Andersen, and H. B. Clausen (2002), Extracting the annual signal from Greenland ice-core chemistry and isotopic records, *Ann Glaciol*, *35*, 131-135.

Rasmussen, S. O., et al. (2006), A new Greenland ice core chronology for the last glacial termination, *J Geophys Res-Atmos*, *111*(D6), -.

Rothlisberger, R., M. Bigler, M. Hutterli, S. Sommer, B. Stauffer, H. G. Junghans, and D. Wagenbach (2000), Technique for continuous high-resolution analysis of trace substances in firn and ice cores, *Environ Sci Technol*, *34*(2), 338-342.

Rupf, I., and G. Radons (2004), New approaches for automated data processing of annually laminated sediments, *Nonlinear Processes in Geophysics*, *11*(5-6), 599-607.

Russell, M. J., and W. J. Holmes (1997), Linear trajectory segmental HMM's, *Ieee Signal Proc Let*, *4*(3), 72-74.

Schmidler, S. C., J. S. Liu, and D. L. Brutlag (2000), Bayesian segmentation of protein secondary structure, *J Comput Biol*, *7*(1-2), 233-248.

Shimada, W., and T. Hondoh (2004), In situ observation of the transformation from air bubbles to air clathrate hydrate crystals using a Mizuho ice core, *J Cryst Growth*, *265*(1-2), 309-317.

Shimohara, K., Miyamoto, A., Hyakutake, K., Shoji, H., Takata, M., Kipfstuhl, S. (2003), Cloudy band observations for annual layer counting on the GRIP and NGRIP, Greenland, deep ice core samples, edited.

Smith, C. L., I. J. Fairchild, C. Spotl, S. Frisia, A. Borsato, S. G. Moreton, and P. M. Wynn (2009), Chronology building using objective identification of annual signals in trace element profiles of stalagmites, *Quaternary Geochronology*, *4*(1), 11-21.

Snyder, D. L., and D. G. Politte (1983), Image-Reconstruction from List-Mode Data in an Emission Tomography System Having Time-of-Flight Measurements, *Ieee T Nucl Sci*, *30*(3), 1843-1849.

Steffensen, J. P., et al. (2008), High-resolution Greenland Ice Core data show abrupt climate change happens in few years, *Science*, *321*(5889), 680-684.

Svensson, A., S. W. Nielsen, S. Kipfstuhl, S. J. Johnsen, J. P. Steffensen, M. Bigler, U. Ruth, and R. Rothlisberger (2005), Visual stratigraphy of the North Greenland Ice Core Project (NorthGRIP) ice core during the last glacial period, *J Geophys Res-Atmos*, *110*(D2), -.

Svensson, A., M. Bigler, E. Kettner, D. Dahl-Jensen, S. J. Johnsen, J. Kipfstuhl, M. Nielsen, and J. P. Steffensen (Submitted 2011), Annual layering in the NGRIP ice core during the Eemian, *Climate of the Past, Discussions*, *7*, 749-773.

Svensson, A., et al. (2006), The Greenland Ice Core Chronology 2005, 15-42 ka. Part 2: comparison to other records, *Quaternary Sci Rev*, *25*(23-24), 3258-3267.

Svensson, A., et al. (2008), A 60 000 year Greenland stratigraphic ice core chronology, *Clim Past*, *4*(1), 47-57.

Takata, M., Y. Iizuka, T. Hondoh, S. Fujita, Y. Fujii, and H. Shoji (2004), Stratigraphic analysis of Dome Fuji Antarctic ice core using an optical scanner, *Annals of Glaciology, Vol 39, 2005*, *39*, 467-472.

Tarantola, A. (2005), *Inverse problem theory and methods for model parameter estimation*, xii, 342 p. pp., Society for Industrial and Applied Mathematics, Philadelphia, PA.

Taylor, J. R. (1997), *An introduction to error analysis : the study of uncertainties in physical measurements*, 2nd ed., xvii, 327 p. pp., University Science Books, Sausalito, Calif.

Taylor, K. C., C. U. Hammer, R. B. Alley, H. B. Clausen, D. Dahljensen, A. J. Gow, N. S. Gundestrup, J. Kipfstuhl, J. C. Moore, and E. D. Waddington (1993), Electrical-Conductivity Measurements from the Gisp2 and Grip Greenland Ice Cores, *Nature*, *366*(6455), 549-552.

Thoraval, L., G. Carrault, and J. J. Bellanger (1994), Heart Signal Recognition by Hidden Markov-Models - the Ecg Case, *Methods of Information in Medicine*, *33*(1), 10-14.

Vinther, B. M., et al. (2006), A synchronized dating of three Greenland ice cores throughout the Holocene, *J Geophys Res-Atmos*, *111*(D13), -.

Viterbi, A. J. (1967), Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm, *Ieee T Inform Theory*, *It13*(2), 260-+.

Wang, Y. J., H. Cheng, R. L. Edwards, Z. S. An, J. Y. Wu, C. C. Shen, and J. A. Dorale (2001), A high-resolution absolute-dated Late Pleistocene monsoon record from Hulu Cave, China, *Science*, *294*(5550), 2345-2348.

Welch, L. R. (2003), The Shannon Lecture: Hidden Markov Models and the Baum-Welch Algorithm, *IEEE Information Theory Society Newsletter*, *53*(4).

Wu, C. F. J. (1983), On the Convergence Properties of the Em Algorithm, *Annals of Statistics*, *11*(1), 95-103.

Yu, S. Z. (2010), Hidden semi-Markov models, *Artificial Intelligence*, *174*(2), 215-243.

Zabin, S. M., and H. V. Poor (1991), Efficient Estimation of Class-a Noise Parameters Via the Em Algorithm, *Ieee T Inform Theory*, *37*(1), 60-72.

Appendix

A1. Nomenclature

Stochastic variable	*Outcome*	*Definition*
	$t \in \{1,2,\dots,T\}$	Indexing number
	$\boldsymbol{o}_t$	Observation(s) at t
	$d \in \mathcal{D} = \{1,2,..,D\}$	Duration of state
S_t	$\ell_j \in \mathcal{L}, j \in \{1,2,\dots,J\}$	State of system (at t)
Q_t	$q_{j,d} = (\ell_j, d)$	Generalized state of system (at t)
Probabilities:	$P(S_{[t_1:t_2]} = \ell_j)$	Layer j starts at t_1 and ends at t_2
	$P(S_{[t_1} = \ell_j)$	Layer j starts at t_1
	$P(S_{t_2]} = \ell_j\ )$	Layer j ends at t_2
	$P(S_t = \ell_j)$	$\boldsymbol{o}_t$ is a part of layer j
Parameters:	θ	Collection of model parameters
	Θ	Collection of hyper-parameters

Sequences are written as e.g. $\boldsymbol{o}_{t_1:t_2}$, this being the observation sequence from t_1 to t_2, and $S_{t_1:t_2}$ being the corresponding sequence of state variables. As a special case, a realization of such a sequence of states of the system is written $s_{t_1:t_2}$.

Probability measures	
$\boldsymbol{a_{(i,d')(j,d)}}$	$P\left(S_{[t+1:t+d]} = \ell_j \middle\vert S_{[t-d'+1:t]} = \ell_i\right)$
$\boldsymbol{a_{ij}}$	$P\left(S_{[t+1} = \ell_j \middle\vert S_{t]} = \ell_i\right)$
$\boldsymbol{b_j(o_{t+1:t+d})}$	$P\left(\boldsymbol{o}_{t+1:t+d} \mid S_{[t+1:t+d]} = \ell_j\right)$
$\boldsymbol{\alpha_t(j,d)}$	$P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right)$
$\boldsymbol{\tilde{\alpha}_t(j)}$	$P\left(S_{t]} = \ell_j, \boldsymbol{o}_{1:t}\right)$
$\boldsymbol{\pi_t(j)}$	$\tilde{\alpha}_{t\leq 0}(j)$
$\boldsymbol{\beta_t(j,d)}$	$P\left(\boldsymbol{o}_{t+1:T} \middle\vert S_{[t-d+1:t]} = \ell_j\right)$
$\boldsymbol{\beta_t}$	$\sum_{\ell_j \in \mathcal{L}} P\left(\boldsymbol{o}_{t+1:T} \middle\vert S_{t]} = \ell_j\right)$
$\boldsymbol{\eta_t(j,d)}$	$P\left(S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:T} \middle\vert \theta\right)$
$\boldsymbol{\overline{\eta}_t(j,d)}$	$P\left(S_{[t-d+1:t]} = \ell_j \middle\vert \boldsymbol{o}_{1:T}, \theta\right)$
$\boldsymbol{\tilde{\eta}_t(j)}$	$P\left(S_{t]} = \ell_j, \boldsymbol{o}_{1:T}\right)$
$\boldsymbol{\gamma_t(j)}$	$P(S_t = \ell_j, \boldsymbol{o}_{1:T})$
$\boldsymbol{\overline{\gamma}_t(j)}$	$P(S_t = \ell_j \vert \boldsymbol{o}_{1:T})$
$\boldsymbol{\ell_{MAP}(t)}$	$\underset{\ell_j}{\text{argmax}} \left\{P(S_t = \ell_j, \boldsymbol{o}_{1:T})\right\}$
$\boldsymbol{\delta_t(j,d)}$	$\max_{s_{1:t-d}} P\left(s_{1:t-d}, S_{[t-d+1:t]} = \ell_j, \boldsymbol{o}_{1:t}\right)$
$\boldsymbol{\breve{\delta}_t(j)}$	$\max_{s_{1:t-1}} P\left(s_{1:t-1}, S_{t]} = \ell_j, \boldsymbol{o}_{1:t}\right)$
$\boldsymbol{\psi_t(j,d)}, \boldsymbol{\breve{\psi}_t(j)}$	Backtracking variables

Layer parameters	
$\boldsymbol{O_j}$	Observation segment corresponding to layer *j*
$\mathbf{X}$	Design matrix giving the layer template
$\boldsymbol{\beta_j}$	Waveform parameter for layer *j*, $\boldsymbol{\beta}_j = \boldsymbol{\varphi} + \boldsymbol{r}_j, \boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\varphi}, \Phi)$
$\boldsymbol{\varphi}$	Mean layer trajectory parameter
$\boldsymbol{r_j} \in \boldsymbol{\mathcal{R}}$	Random effect vector for layer *j*
$\boldsymbol{\Phi}$	Variance of random effect vectors
$\boldsymbol{E_j}$	Gaussian white noise vector for layer *j*
$\boldsymbol{\sigma_\varepsilon^2}$	Variance of white noise
$\boldsymbol{\mu_d}, \boldsymbol{\sigma_d}$	Layer thickness distribution parameters

A2. The lognormal distribution

A random variable, X, whose logarithm is normally distributed, i.e. $Y = \log X \sim \mathcal{N}(\mu, \sigma^2)$, is said to follow a lognormal probability distribution. It can be written as:

$$X \sim Log\ \mathcal{N}(\mu, \sigma^2)$$

The parameters μ and σ are sometimes termed the "location parameter" and "scale parameter".

The base of the lognormal transformation can be chosen freely. In *Andersen et al.* [2006] $\log_{10}$ was used to describe the annual layer thickness probability distribution, while for this work the natural logarithm, $\log_e$, was chosen. The transformation between the two, however, is simple:

$$\mu_e = \log_e(10)\mu_{10}$$

$$\sigma_e = \log_e(10)\sigma_{10}$$

Under multiplication with a constant, the location parameter of the lognormal distribution changes, but the scale parameter does not. This is e.g. the case when using a simple flow-model to correct for the strain-induced thinning of the annual layers with depth, in which case we have for the corrected annual layer thicknesses:

$$\lambda_{corr} = k \cdot \lambda$$

$$\log \lambda_{corr} = \log(k \cdot \lambda) = \log \lambda + \log k \ \sim \mathcal{N}(\mu, \sigma^2) + k = \mathcal{N}(\mu + k, \sigma^2)$$

Hence, under the assumption of constant scale parameter of the annual accumulation rates over time, also the scale parameter of the annual layer thicknesses should remain the same for all depths.

Similar goes for the transformation of λ between measurements in different units. Assuming the annual layer thicknesses to be given in cm, and having a resulting lognormal distribution with parameters $\log \lambda_{cm} \sim N(\mu_{cm}, \sigma^2)$, the resulting distribution in m can be computed as:

$$\lambda_m = \lambda_{cm} \cdot 10^{-2}$$

$$\log \lambda_m = \log(\lambda_{cm} \cdot 10^{-2}) = \log \lambda_{cm} + \log(10^{-2})$$

That is, the resulting distribution is given by:

$$\log \lambda_m \sim N(\mu_{cm} + \log 10^{-2}, \sigma^2)$$

Note, that although the location parameter changes, the scale parameter does not.

The mean value, median and mode of the lognormal distribution can be computed as:

$$mean = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$median = \exp(\mu)$$

$$mode = \exp(\mu - \sigma^2)$$

A3. Convergence of the EM-algorithm

According to the theory behind the EM-algorithm, the likelihood of the parameter values will under most circumstances converge to a (local) maximum when repeatedly iterating between the expectation step (E-step) and the maximization step (M-step). Following the derivation in *Gupta and Chen* [2011], a proof will here be given that the likelihood of the parameters will indeed always monotonously increase during these iterations.

The log-likelihood function of a parameter θ is defined as:

$$\log L(\theta|\boldsymbol{o}_{1:T}) = \log P(\boldsymbol{o}_{1:T}|\theta)$$

Treating the full hidden state sequence $s_{1:T} \in \mathcal{L}^T$ as unknown, the above can be rewritten as the sum of the joint probability of $\boldsymbol{o}_{1:T}$ and $s_{1:T}$ when summed over all possible realizations of $s_{1:T}$:

$$\log L(\theta|\boldsymbol{o}_{1:T}) = \log \sum_{s_{1:T}\in\mathcal{L}^T} P(\boldsymbol{o}_{1:T}, s_{1:T}|\theta)$$

In this expression, both numerator and denominator can be multiplied with $P(s_{1:T}|\theta^{(k)}, \boldsymbol{o}_{1:T})$, hereby allowing the sum to be rewritten as an expectation:

$$\begin{aligned}\log L(\theta|\boldsymbol{o}_{1:T}) &= \log \sum_{s_{1:T}\in\mathcal{L}^T} \frac{P(\boldsymbol{o}_{1:T}, s_{1:T}|\theta)}{P(s_{1:T}|\theta^{(k)}, \boldsymbol{o}_{1:T})} P(s_{1:T}|\theta^{(k)}, \boldsymbol{o}_{1:T}) \\ &= \log \mathbb{E}\left[\frac{P(\boldsymbol{o}_{1:T}, s_{1:T}|\theta)}{P(s_{1:T}|\theta^{(k)}, \boldsymbol{o}_{1:T})} \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\right]\end{aligned}$$

By Jensen's inequality (see e.g. *Bishop* [2006]), it must then hold that:

$$\log L(\theta|\boldsymbol{o}_{1:T}) \geq \mathbb{E}\left[\log\left(\frac{P(\boldsymbol{o}_{1:T}, s_{1:T}|\theta)}{P(s_{1:T}|\theta^{(k)}, \boldsymbol{o}_{1:T})}\right) \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\right] \tag{3.1}$$

Now, Bayes' theorem will be utilized, along with the knowledge that the probability of the observation sequence is fully determined by the underlying state sequence, i.e.:

$$P(\boldsymbol{o}_{1:T}, s_{1:T}|\theta) = P(s_{1:T}|\theta)P(\boldsymbol{o}_{1:T}|s_{1:T}, \theta) = P(s_{1:T}|\theta)P(\boldsymbol{o}_{1:T}|s_{1:T})$$

$$P\left(s_{1:T}\middle|\theta^{(k)}, \boldsymbol{o}_{1:T}\right) = \frac{P(s_{1:T}, \boldsymbol{o}_{1:T}|\theta^{(k)})}{P(\boldsymbol{o}_{1:T}|\theta^{(k)})} = \frac{P\left(s_{1:T}\middle|\theta^{(k)}\right)P(\boldsymbol{o}_{1:T}|s_{1:T})}{P(\boldsymbol{o}_{1:T}|\theta^{(k)})}$$

Inserting these in (3.1), and eliminating the common factor of $P(\boldsymbol{o}_{1:T}|s_{1:T})$, we arrive at the following expression:

$$\begin{aligned}\log L(\theta|\boldsymbol{o}_{1:T}) &\geq \mathbb{E}\left[\log\left(\frac{P(s_{1:T}|\theta)}{P(s_{1:T}|\theta^{(k)})/P(\boldsymbol{o}_{1:T}|\theta^{(k)})}\right) \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\right] \\ &= \mathbb{E}\left[\log\left(\frac{P(s_{1:T}|\theta)P\left(\boldsymbol{o}_{1:T}\middle|\theta^{(k)}\right)}{P\left(s_{1:T}\middle|\theta^{(k)}\right)}\right) \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\right] \\ &= \mathbb{E}\left[\log P(s_{1:T}|\theta) \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\right] + \log P\left(\boldsymbol{o}_{1:T}\middle|\theta^{(k)}\right) \\ &\quad - \mathbb{E}\left[\log P\left(s_{1:T}\middle|\theta^{(k)}\right) \middle| \theta^{(k)}, \boldsymbol{o}_{1:T}\right]\end{aligned}$$

Recall that the Q-function appearing in the EM-algorithm is given by:

$$\begin{aligned} Q(\theta|\theta^{(k)}) &\equiv \mathbb{E}\big[\log L(\theta|s_{1:T}, \boldsymbol{o}_{1:T}) \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\big] \\ &= \mathbb{E}\big[\log P(s_{1:T}, \boldsymbol{o}_{1:T}|\theta) \mid \theta^{(k)}, \boldsymbol{o}_{1:T}\big] \end{aligned}$$

Inserting this definition of the Q-function in the equation above yields:

$$\begin{aligned} \log L(\theta|\boldsymbol{o}_{1:T}) &\geq Q\big(\theta\big|\theta^{(k)}\big) + \log P\big(\boldsymbol{o}_{1:T}\big|\theta^{(k)}\big) - Q\big(\theta^{(k)}\big|\theta^{(k)}\big) \\ &= \log L(\theta^{(k)}|\boldsymbol{o}_{1:T}) + Q\big(\theta\big|\theta^{(k)}\big) - Q\big(\theta^{(k)}\big|\theta^{(k)}\big) \end{aligned} \tag{3.2}$$

At each M-step in the iteration of the EM-algorithm, the value of θ is found, for which the function $Q(\theta|\theta^{(k)})$ is maximized, and hence, it must necessarily hold that $Q\big(\theta\big|\theta^{(k)}\big) \geq Q\big(\theta^{(k)}|\theta^{(k)}\big)$. According to the inequality (3.2), this in turn implies that $\log L(\theta|\boldsymbol{o}_{1:T}) \geq \log L\big(\theta^{(k)}\big|\boldsymbol{o}_{1:T}\big)$. Consequently, the likelihood of the new set of model parameters estimated during the M-step can never be lower than that of the original parameters.

The same is true for a MAP estimate of the parameter values. In this case, the function to be maximized at each M-step is given by:

$$R\big(\theta|\theta^{(k)}\big) = Q\big(\theta\big|\theta^{(k)}\big) + \log P(\theta)$$

Adding $\log P(\theta)$ to both sides of (3.2), and adding and subtracting $\log P(\theta^{(k)})$ on the right side of the inequality, yields:

$$\begin{aligned} \log L(\theta|\boldsymbol{o}_{1:T}) + \log P(\theta) &\geq \log L\big(\theta^{(k)}|\boldsymbol{o}_{1:T}\big) + Q\big(\theta\big|\theta^{(k)}\big) - Q\big(\theta^{(k)}\big|\theta^{(k)}\big) + \log P(\theta) \\ &= \log L(\theta^{(k)}|\boldsymbol{o}_{1:T}) + \log P(\theta^{(k)}) + \big(Q\big(\theta\big|\theta^{(k)}\big) + \log P(\theta)\big) \\ &\quad - \big(Q\big(\theta^{(k)}\big|\theta^{(k)}\big) + \log P(\theta^{(k)})\big) \end{aligned}$$

By maximizing $R\big(\theta\big|\theta^{(k)}\big)$ at each M-step, it is ensured that:

$$Q\big(\theta\big|\theta^{(k)}\big) + \log P(\theta) = R\big(\theta\big|\theta^{(k)}\big) \geq R\big(\theta^{(k)}\big|\theta^{(k)}\big) = Q\big(\theta^{(k)}\big|\theta^{(k)}\big) + \log P\big(\theta^{(k)}\big)$$

Hence, it must also hold that:

$$\log L(\theta|\boldsymbol{o}_{1:T}) + \log P(\theta) \geq \log L(\theta^{(k)}|\boldsymbol{o}_{1:T}) + \log P(\theta^{(k)})$$

This is exactly what we wanted to know, as it implies that the posterior probability of the parameter values will never decrease:

$$\log P(\boldsymbol{o}_{1:T}|\theta) \geq \log P(\boldsymbol{o}_{1:T}|\theta^{(k)})$$

In this way, the EM-algorithm presents a method in which the (posterior) most likely set of parameter values can be estimated, and it does so without requiring any knowledge on the hidden state sequence giving rise to the observed data. All that is needed is a way to calculate the function $Q(\theta|\theta^{(k)})$ based on observations (which in our case is provided by either the Forward-Backward or the Viterbi algorithm) and an initial guess of model parameter values. Iteratively calculating (the E-step) and maximizing (the M-step) the Q-function, the likelihood of the chosen parameter values will either stay at the same level or increase.

The EM-algorithm provides a general procedure, which can be used for solving a range of maximization problems. However, no guarantee is given that the obtained maximum is a global maximum. Caused by the deterministic behavior of the algorithm, it might get caught up in a local maximum of the likelihood function if given a bad initial estimate and a complex likelihood function. Multiple random starts may be used to better ensure that this is not the case.

One of the weaknesses of the EM-algorithm is a relatively slow convergence, in particular in case of models with many unknown parameters. To resolve this problem, various extensions of the EM-algorithm have been suggested. A common extension is to amend the algorithm to include intermediate steps of conditional and/or constrained maximizations, which can help to speed up its convergence [*Bishop*, 2006; *Kim and Smyth*, 2006; *Meng and Rubin*, 1993][1].

[1] Meng, X.-L., and D. B. Rubin (1993), Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika*, *80*(2), 267-278.

A4. EM update equations

A4.1 The differential of residuals

In section 5.3.3, it was stated that the following identity holds:

$$\frac{\partial}{\partial \boldsymbol{\varphi}}\left(\left(\boldsymbol{O}_j - \mathrm{X}(\boldsymbol{\varphi} + \boldsymbol{r}_j)\right)^{\mathsf{T}} \mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}(\boldsymbol{\varphi} + \boldsymbol{r}_j)\right)\right) = -2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}(\boldsymbol{\varphi} + \boldsymbol{r}_j)\right)$$

Proof:

Let's first re-write the equation using the notation $\boldsymbol{a} \equiv \boldsymbol{O}_j - \mathrm{X}\boldsymbol{r}_j$. With this abbreviation, the differential is given as:

$$\frac{\partial}{\partial \boldsymbol{\varphi}}((\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi})^{\mathsf{T}}\,\mathrm{W}^{-1}(\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi}))$$

By completing the squares, we find:

$$\frac{\partial}{\partial \boldsymbol{\varphi}}(\boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\boldsymbol{a} - \boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi} - \boldsymbol{\varphi}^{\mathsf{T}}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\boldsymbol{a} + \boldsymbol{\varphi}^{\mathsf{T}}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi})$$

Realizing that $\boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi}$ is just a number, we have that $\boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi} = (\boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi})^{\mathsf{T}} = \boldsymbol{\varphi}^{\mathsf{T}}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\boldsymbol{a}$. It has here been used that W is a diagonal matrix, implying that W^{-1} is too, and therefore $(\mathrm{W}^{-1})^{\mathsf{T}} = \mathrm{W}^{-1}$. Consequently, the equation above can be reduced to:

$$\frac{\partial}{\partial \boldsymbol{\varphi}}((\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi})^{\mathsf{T}}\,\mathrm{W}^{-1}(\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi})) = -2\frac{\partial \boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi}}{\partial \boldsymbol{\varphi}} + \frac{\partial \boldsymbol{\varphi}^{\mathsf{T}}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi}}{\partial \boldsymbol{\varphi}}$$

Consider each part separately. As it holds for any vector $\boldsymbol{v}$, that $\frac{\partial}{\partial \boldsymbol{\varphi}}(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{\varphi}) = \boldsymbol{v}$ [*Petersen and Pedersen*, 2008], we get for the first part:

$$\frac{\partial \boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi}}{\partial \boldsymbol{\varphi}} = (\boldsymbol{a}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X})^{\mathsf{T}} = \mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\boldsymbol{a}$$

For the second part, we can utilize the fact that it holds for any matrix Y that $\frac{\partial}{\partial \boldsymbol{\varphi}}\boldsymbol{\varphi}^{\mathsf{T}}\mathrm{Y}\boldsymbol{\varphi} = (\mathrm{Y} + \mathrm{Y}^{\mathsf{T}})\boldsymbol{\varphi}$ [*Petersen and Pedersen*, 2008]:

$$\frac{\partial \boldsymbol{\varphi}^{\mathsf{T}}\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi}}{\partial \boldsymbol{\varphi}} = (\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X} + (\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X})^{\mathsf{T}})\boldsymbol{\varphi} = 2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi}$$

Inserting these, and finally replacing with the original expression for $\boldsymbol{a}$, provides the desired result:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\varphi}}\left((\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi})^{\mathsf{T}}\,\mathrm{W}^{-1}(\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi})\right) &= -2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\boldsymbol{a} + 2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{\varphi} = -2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}(\boldsymbol{a} - \mathrm{X}\boldsymbol{\varphi}) \\ &= -2\mathrm{X}^{\mathsf{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}(\boldsymbol{\varphi} + \boldsymbol{r}_j)\right)\end{aligned}$$

A4.2 **Expectation value of weighted residuals**

In section 5.3.5, the following identity was put forward without proof:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{E}_j^{\mathrm{T}}\mathrm{W}\boldsymbol{E}_j|\,\boldsymbol{O}_j,\theta^{(k)}\right] &\equiv \int_{\boldsymbol{r}_j} p\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\left(\boldsymbol{O}_j-\mathrm{X}(\boldsymbol{\varphi}+\boldsymbol{r}_j)\right)^{\mathrm{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j-\mathrm{X}(\boldsymbol{\varphi}+\boldsymbol{r}_j)\right)d\boldsymbol{r}_j\\ &=\left(\boldsymbol{O}_j-\mathrm{X}(\boldsymbol{\varphi}+\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}])\right)^{\mathrm{T}}\mathrm{W}^{-1}\left(\boldsymbol{O}_j-\mathrm{X}(\boldsymbol{\varphi}+\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}])\right)\\ &\quad+\mathrm{tr}\left(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)\end{aligned}$$

Proof:

To simplify the notation, a vector defined as $\boldsymbol{b}\equiv\boldsymbol{O}_j-\mathrm{X}\boldsymbol{\varphi}$ will be used in the following. The integral above can then be stated as:

$$\begin{aligned}&\mathbb{E}\left[\boldsymbol{E}_j^{\mathrm{T}}\mathrm{W}\boldsymbol{E}_j|\,\boldsymbol{O}_j,\theta^{(k)}\right]=\int_{\boldsymbol{r}_j} p\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\left(\boldsymbol{b}-\mathrm{X}\boldsymbol{r}_j\right)^{\mathrm{T}}\mathrm{W}^{-1}\left(\boldsymbol{b}-\mathrm{X}\boldsymbol{r}_j\right)d\boldsymbol{r}_j\\ &=\int_{\boldsymbol{r}_j} p\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\left(\boldsymbol{b}^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{b}-\boldsymbol{b}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j-\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{b}+\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j\right)d\boldsymbol{r}_j\\ &=\boldsymbol{b}^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{b}-\boldsymbol{b}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\int_{\boldsymbol{r}_j} p\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\boldsymbol{r}_j d\boldsymbol{r}_j-\int_{\boldsymbol{r}_j} p\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\boldsymbol{r}_j^{\mathrm{T}}d\boldsymbol{r}_j\,\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{b}\\ &\qquad+\int_{\boldsymbol{r}_j} p\left(\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right)\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j d\boldsymbol{r}_j\\ &=\boldsymbol{b}^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{b}-\boldsymbol{b}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\,\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]-\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\boldsymbol{b}\\ &\qquad+\mathbb{E}\left[\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\\ &=\left(\boldsymbol{b}-\mathrm{X}\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)^{\mathrm{T}}\mathrm{W}^{-1}\left(\boldsymbol{b}-\mathrm{X}\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)\\ &\qquad-\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\,\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]+\mathbb{E}\left[\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\end{aligned}$$

Consider the last term only: As $\mathbb{E}\left[\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$ is just a number, treating it like a matrix and taking its trace will not change anything. This subsequently allows us to change the sequence of the variables, as $\mathrm{tr}(\mathrm{ABC})=\mathrm{tr}(\mathrm{BCA})=\mathrm{tr}(\mathrm{CAB})$ for any three matrices A, B and C. We get:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]&=\mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j\right)|\boldsymbol{O}_j,\theta^{(k)}\right]\\ &=\mathbb{E}\left[\mathrm{tr}\left(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}\right)|\boldsymbol{O}_j,\theta^{(k)}\right]\\ &=\mathrm{tr}\left(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\,\mathbb{E}\left[\boldsymbol{r}_j\boldsymbol{r}_j^{\mathrm{T}}|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\end{aligned}$$

For a vector $\boldsymbol{x}$, which is normally distributed with mean $\boldsymbol{\mu}$ and covariance Σ, we have that $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]=\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+\Sigma$. The above can therefore be re-written in terms of the conditional mean and covariance of $\boldsymbol{r}_j$, $\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$ and $\mathrm{cov}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]$:

$$\begin{aligned}&\mathbb{E}\left[\boldsymbol{r}_j^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\\ &\qquad=\mathrm{tr}\left(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\left(\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]^{\mathrm{T}}+\mathrm{cov}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right)\right)\end{aligned}$$

$$= \mathrm{tr}\left(\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]^{\intercal}\right) + \mathrm{tr}\left(\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$
$$= \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]^{\intercal}\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]\right) + \mathrm{tr}\left(\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$
$$= \mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right]^{\intercal}\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathbb{E}\left[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}\right] + \mathrm{tr}\left(\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$

Comparing this expression to the remaining terms in the equation for $\mathbb{E}\left[\boldsymbol{E}_j^{\intercal}\mathrm{W}\boldsymbol{E}_j|\,\boldsymbol{O}_j,\theta^{(k)}\right]$, the similarities between the two last terms can be seen. We hence arrive at the following expression for the expectation value of the weighted squared residuals:

$$\mathbb{E}\left[\boldsymbol{E}_j^{\intercal}\mathrm{W}\boldsymbol{E}_j|\,\boldsymbol{O}_j,\theta^{(k)}\right]$$
$$= \left(\boldsymbol{b} - \mathrm{X}\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)^{\intercal}\mathrm{W}^{-1}\left(\boldsymbol{b} - \mathrm{X}\mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)$$
$$+ \mathrm{tr}\left(\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$

Or, by insertion of the original vector $\boldsymbol{b} = \boldsymbol{O}_j - \mathrm{X}\boldsymbol{\varphi}$**:**

$$\mathbb{E}\left[\boldsymbol{E}_j^{\intercal}\mathrm{W}\boldsymbol{E}_j|\,\boldsymbol{O}_j,\theta^{(k)}\right]$$
$$= \left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi} + \mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)\right)^{\intercal}\mathrm{W}^{-1}\left(\boldsymbol{O}_j - \mathrm{X}\left(\boldsymbol{\varphi} + \mathbb{E}[\boldsymbol{r}_j|\boldsymbol{O}_j,\theta^{(k)}]\right)\right)$$
$$+ \mathrm{tr}\left(\mathrm{X}^{\intercal}\mathrm{W}^{-1}\mathrm{X}\,\mathrm{cov}\left[\boldsymbol{r}_j\middle|\boldsymbol{O}_j,\theta^{(k)}\right]\right)$$