USING EXPERT KNOWLEDGE WHEN THE DATA MODEL IS UNKNOWN WITH AN APPLICATION IN MODELING THE MIXED LAYER OF THE ATLANTIC OCEAN

by

Ana Grohovac Rappold

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Prof. Michael Lavine, Supervisor

_____
Prof. Susan Lozier

_____
Assistant Prof. Feng Liang

_____
J.B. Duke Prof. Alan Gelfand

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2004

ABSTRACT

(Statistics)

USING EXPERT KNOWLEDGE WHEN THE DATA MODEL IS UNKNOWN WITH AN APPLICATION IN MODELING THE MIXED LAYER OF THE ATLANTIC OCEAN

by

Ana Grohovac Rappold

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Prof. Michael Lavine, Supervisor

_____
Prof. Susan Lozier

_____
Assistant Prof. Feng Liang

_____
J.B. Duke Prof. Alan Gelfand

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2004

Abstract

Oceanographers are interested in studying decadal variability of the ocean's heat content through the depth of the Mixed Layer M. This is the top layer of the ocean where the waters come together through turbulent mixing and convection, which creates a nearly uniform temperature layer in the ocean. It evolves with the season's annual cycle of surface temperatures and is smoothly varying in spaces. Understanding the sampling distribution of the data is difficult because of the spatio-temporal dependencies, because M does not uniquely determine the distribution of temperatures, and because the estimator $\hat{M}$ is not known. However, before and after observing the temperature profile, an expert has a prior and posterior belief. The novel idea of this thesis is how to translate the expert's uncertainty into a likelihood function.

We elicit the expert's posterior belief on a partition in the parameter space such that her prior probabilities may be considered constant across all elements of the partition. For any single profile, the expert's posterior belief is expressed by the function $h(M|data)$. To model spatio-temporal dependencies, we suggest substituting the likelihood by $h(M)$. The function h is not a product of a known density function, but it may be viewed as a distribution of the pivotal variables and is independent of the data rather than M, as usually defined. The pivotal variables have a strong scientific interpretation, and they serve as a transformation between the observed space and the transformed space. The posterior inference is given directly from the distribution of the transformed variables. The resulting posterior distribution is closely associated with *direct probabilities* of Dempster and *fiducial distributions* of Fisher. In general, there need not exist a marginal distribution for M which yields the posterior inference identical to the one arising from the pivotal quantities, but

there always exists a likelihood function which, by the fiducial argument, gives the same posterior inference.

We also provide the inference using traditional methods when the sampling model is unknown. Although modeling known features in the data may provide valuable results, the posterior distribution of M for any single profile may not agree with the expert opinion. We provide a discussion as to why this is true. In the final chapter, we propose a flexible spatio-temporal model to help us understand the nature of the long-term changes in the depth of the Mixed Layer.

Acknowledgements

I would like to thank my academic advisors, Professors Michael Lavine and Susan Lozier, for their constant support during the development of this thesis. They have always provided me with encouragement and the most valuable discussions.

I would also like to thank the other committee members, Prof. Alan Gelfand and Assistant Professor Feng Liang. Discussions with Prof. Liang has been fruitful, and I felt privileged to have her as a part of my committee. Professor Gelfand has always painted a positive picture for me and all other students at ISDS. I thank both of them for being supportive of my work and providing advice along the way.

I would also like to thank other members of the faculty who have, in many ways, helped me stay on track: Paul Marriott, Peter Muller, and Val Johnson, all of whom spent time with us at ISDS. Our system administrator, Eric Van Gyzen, has graciously pulled me out of many computing struggles. We all have been lucky to have him around.

I also want to thank my friends here at ISDS. They have shared in the moments of struggles and successes all the way through. I have made irreplaceable friendships in the last few years.

I want to thank my family on the other side of the Atlantic for providing tremendous support, love, and courage throughout my life. And finally, my warmest thoughts I reserve for my husband Eric. If I was asked to draw a picture of the most loving and thoughtful person in my life, I would draw a picture of him.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The primary focus of environmental research in recent decades has been the detection of anthropogenic impacts on the global climate. In oceanographic research, detection of an anthropogenic signal has been focused on monitoring the heat content and salinity of the ocean waters. The ocean is a critical sink of atmospheric heat. As such, any observed long term trend in the heat content of the ocean, although weak, has an important role in explaining discrepancies between observed and predicted atmospheric temperatures over this past century of increased greenhouse gasses and global warming. Furthermore, changes in the ocean's heat content can alter dynamic ocean processes through complex thermodynamic feedback. In this thesis, we will study the ocean's heat by analyzing the characteristics of the Mixed Layer, which is defined as the top, nearly constant temperature layer of the surface waters formed through wind driven turbulent mixing and convection. Figure 1.1, taken from Mellor (1996), illustrates the evolution of the Mixed Layer depth (M); temperatures are plotted on the horizontal axis, and the Roman numerals mark months of the year, placing the season of warming in the left panel and the season of cooling in the right

panel. The Mixed Layer fepth M is a quantity that varies cyclically within a year due to seasonal temperature changes, in space due to the latitudinal and longitudinal temperature variation, and annually due to climate variability. Our interest is to study the decadal impact of increasing heat stress at the surface on the depth of the Mixed Layer.

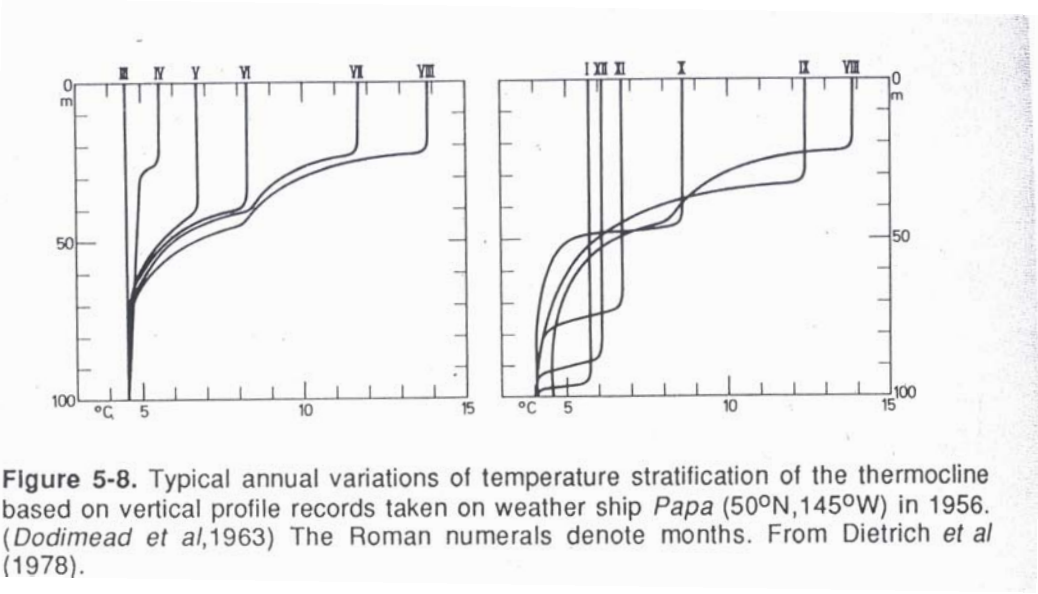


Figure 5-8. Typical annual variations of temperature stratification of the thermocline based on vertical profile records taken on weather ship *Papa* (50°N,145°W) in 1956. (*Dodimead et al*,1963) The Roman numerals denote months. From Dietrich *et al* (1978).

Figure 1.1: Hypothetical annual evolution of a thermal profile.

Available measurements of temperature, salinity, pressure, and oxygen, dated since the 1900s, are stored in archives at the National Oceanographic Data Center(NODC). Data is collected by vessels, research and commercial, leading to obvious restrictions on the spatial and temporal resolution of sampling procedures. Although spatio-temporal analysis of M is necessary, understanding the data generating process is difficult due to the lack of directly comparable data in both time and space. The time scales for ocean processes prohibit treating data collected within a period of several days as independent. On the other hand, samples taken at the same location but in different years cannot be directly compared due to the inter-annual variability of temperatures. These issues, together with the fact that there is no known esti-

mator for M, make it difficult to determine what parametric family is best suited to model the long-term trend in the annual cycle.

The novel idea of this thesis is the elicitation of a *likelihood function* that uses an expert opinion when the data model is not known. We create partition for the parameter of interest and elicit the joint distribution of observables and unobservables in the form of a small number of pivotal variables with strong scientific interpretations. These pivotal variables, as usual, are functions of the data and the parameters but their distribution is regardes as independent of the data rather than of the parameter. Conditioning on the data, pivotal quantities give us the posterior conditional distribution on M. Because the pivotal variables are elicited from the expert's opinion, we believe they yield the correct posterior uncertainty in M. Another name for the elicited posteriors is the *fiducial* distribution: a likelihood function seen as a posterior density with a constant measure on θ over all partitions. Therefore, we set the likelihood as a function of M to be the functional form that approximates the posterior, *i.e.* the distribution of the pivotal quantities. We say this likelihood is approximate because it need not be proportional to the product of any probability density function, as usually defined. We discuss a parallel between the elicited likelihood and *direct probabilities* introduced by Dempster (1963). Because the posterior conditional distribution of M is interpreted as a distribution obtained from the pivotal quantities (such as Dempster, 1963), we give a general recipe for obtaining a substitute likelihood function when the data model is not known.

We compare the performance of the algorithmic likelihood to a proper probability model in which we model a monotonically decreasing function of temperatures with depth via transformation as a Dirichlet Process. To "match" the oceanographer's posterior belief with the posterior of the full probability model, it is necessary to elicit a prior precision parameter of the Dirichilet Process. The point is made that when

adopting full probability models without knowing the true sampling distribution, it becomes unclear where the expert opinion belongs: in the prior or in the likelihood. We also show that the algorithimic likelihood is robust and point out weaknesses in the full probability model.

Spatial and temporal dependence among the profiles is modeled at the second stage of hierarchy, and long term trends are additive to the inter-annual spatial Gaussian process, which accounts for a smoothly varying seasonal cycle. The model we propose allows for easy exploration of the long term rates of change; rates by years, by decades, by specific regions, *etc.*. As cruises typically last no more than a few weeks, it is reasonable to assume that observations taken on the same expedition reflect similar variations from the overall spatio-temporal mean and are not regarded as independent. Therefore, we include a "cruise effect" random variable that accounts for small scale systematic deviations. The spatio-inter-annual process M is modeled as a convolution (Higdon, 1998) of a two dimensional white noise process with a separable Gaussian kernel, which results in a latitude-dependent annual cycle. In the temporal domain, the kernel is chosen to have support on a unit circle which ensures periodicity. Construction of a spatio-temporal process via convolution adds computationally attractive features because the dimension of the model is the dimension of the white noise process and not that of the observations. The size of the data set imposes serious restrictions on elaborating the model.

As a result of challenges in analyzing oceanographic data, many of the analyses rely on data from deterministic mathematical models. Well-tuned models may provide us with important information. For example, we can study changes in output fields as a function of input values such as heat and wind fluxes. In essence, models can serve as controlled experiments which are validated by observations. It is crucial to establish methods to analyze the observational data for its own value and

for checking the performance of the physical models. Because M is not an observable quantity and its sampling distribution is not known, all the previous analyses have been based on ad-hoc methods which yield arbitrary bias and variability. Our approach is computationally feasible and yields proper probabilistic estimates of uncertainty.

In the final section, we will show the results of our analysis. We are looking at an interesting finding that, overall, in both the subtropical and subpolar gyre of the North Atlantic Ocean, the Mixed Layer has been shoaling. This finding is contradictory to the previous belief that, in the subpolar gyre, warming in the atmosphere would produce cooling of the surface waters, and, consequently, deepening of the Mixed Layer. Our results are supported by further illustration of temperature behavior at different depths. In both regions, we notice significant warming trends of the deeper waters, indicating that the ocean is storing heat deeper than previously believed. Temperature trends at deeper levels, in particular $250\,m$ in the subtropical gyre and $500\,m$ in the subpolar gyre, seem to be better indicators of the trend in M than surface temperature.

Chapter 2

Data

2.1 Description of the data

In this thesis, we will use temperature data from the hydrographic stations stored at the National Oceanographic Data Center (NODC) in Maryland. NODC compiles, stores, and distributes satellite, in situ physical, and chemical data furnished by both US research teams and multinational programs such as the World Ocean Circulation Experiment (WOCE 1992 - present). The hydrographic data used here is collected by research and commercial vessels; physical locations of all observed data is depicted in figure 2.1, while specific cruises that have been repeated over the years through the WOCE program (WHP) are plotted in figure 2.1. Two rectangles define the boundaries of the observations used in the analysis and are data subsets of the subtropical and subpolar gyre.

We are interested in modeling a particular aspect of the thermal profile, namely, the bottom of the Mixed Layer and its long term trend. The Mixed Layer is a concept and is not observable. Temperature profiles are characterized by a constant temperature in the top layer of the ocean and an exponential decay below M. From a statistical point of view, we define two issues:

Figure 2.1: Physical location of the observed data in the Atlantic Ocean.

1. There is no known estimator for M

2. The Mixed Layer depth evolves seasonally and is spatially dependent.

2.1.1 Distribution of the Mixed Layer at a given location

Data from ships consists of water temperatures measured every δ meters of depth in the ocean. While the choice of δ is part of an experimental design and varies among different expeditions, it is typically about 5 to 10 meters. Currently, the estimator of the Mixed Layer depth, $\hat{M}$, is commonly recognized as the minimum depth at which the change of temperature from the surface exceeds $0.2\,^oC$. While the $0.2\,^oC$ rule is completely heuristic, temperatures are linearly interpolated between the observed depths, which creates further systematic bias. The Mixed Layer depth may also be

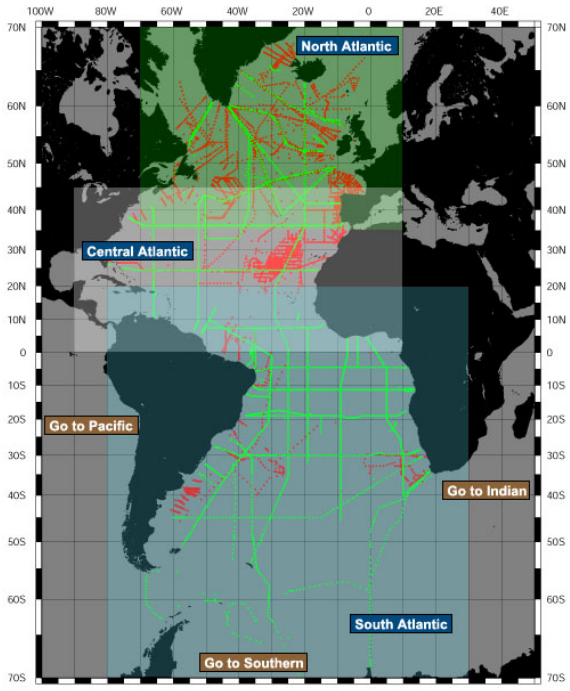


Figure 2.2: WHP repeated cruises.

studied in terms of density rather than temperature, in which case $\hat{M}$ is defined as the minimum depth where a change of 0.125 σ units from the surface is observed. Although also heuristic, this rule seems to perform more robustly than the one based on temperature. The *uncertainty* about M is studied through the variability of $\hat{M}$ using a sequence of rules such as $0.2 \pm 0.01 * j$ for $j = 1, ..., 10$. However, such variability does not represent the uncertainty in M; rather, it describes the sensitivity towards the choice of the threshold rule.

For a valid representation of the uncertainty, we need a full probability model based on the joint distribution of observable and unobservable variables. In general, we don't have repeated samples in spatial and temporal dimensions with oceanographic data. A good start in understanding the data model is to explore the theoretical knowledge about the transport of heat through the ocean water. The relationship between temperature at depth d and surface forcing is fairly complex and

depends on a series of estimated constants such as velocity, kinetic heat flux, etc. Let $x(d), u, v, t$ denote temperature as a function of depth, two horizontal coordinates, and time, respectively. Then under the assumption of horizontal homogeneity and temporal stationarity,

$$\partial x(d)/\partial u = \partial x(d)/\partial v = \partial x(d)/\partial t = 0,$$

oceanographers believe that the temperature profile may be represented through the following form:

$$x(d) \;=\; A + B\,\exp\left\{-\frac{w}{k}\,d\right\}$$

where k denotes the eddy diffusivity coefficient measured in m^2/s and w stands for the velocity measured in m/s. Since none of the coefficients are known exactly and may vary over space and time, oceanographers use approximations.

Constants A and B are further determined for two layers of the ocean separately. Let Q_0 denote kinematic heat flux, then for the upper layer we have:

$$\begin{aligned} x(d) &= x(0) - \frac{Q_0}{w}\left[1 - e^{-\frac{w}{k_1}\,d}\right] \quad \text{for } d < M \\ &\approx x(0) - \frac{Q_0}{k_1}\,d \qquad \text{since } w = 0 \text{ for } d < M \\ &\approx x(0) \end{aligned}$$

Below the mixed layer we have:

$$x(d) \;=\; x(0)\left[1 - \frac{Q_0\,M}{k_1\,t(s)}\right]\,\exp\left\{\frac{-w(d-M)}{k_2}\right\} \quad \text{for } d \geq M \tag{2.1}$$

$$\approx\; x(0)\,\exp\left\{\frac{-w(d-M)}{k_2}\right\} \tag{2.2}$$

Notice that at $d = M$ continuity must hold, and that $t(d) \to 0$ as $d \to \infty$. Ocean temperatures actually asymptote around 2.1^oC, but this does not present a problem since all temperatures may be rescaled relative to the ocean floor temperatures.

In figure 2.3 we plotted an observed profile where M is clearly identified between the 5th and 6th observed depth, or $51m$ to $61m$, and overlayed a plot of theoretical approximations for the range of realistic values of k_2/w (Mellor, 1996, Ch. 2 and 5). Note the discrepancies between the theoretical and observed profiles in figure 2.3; the

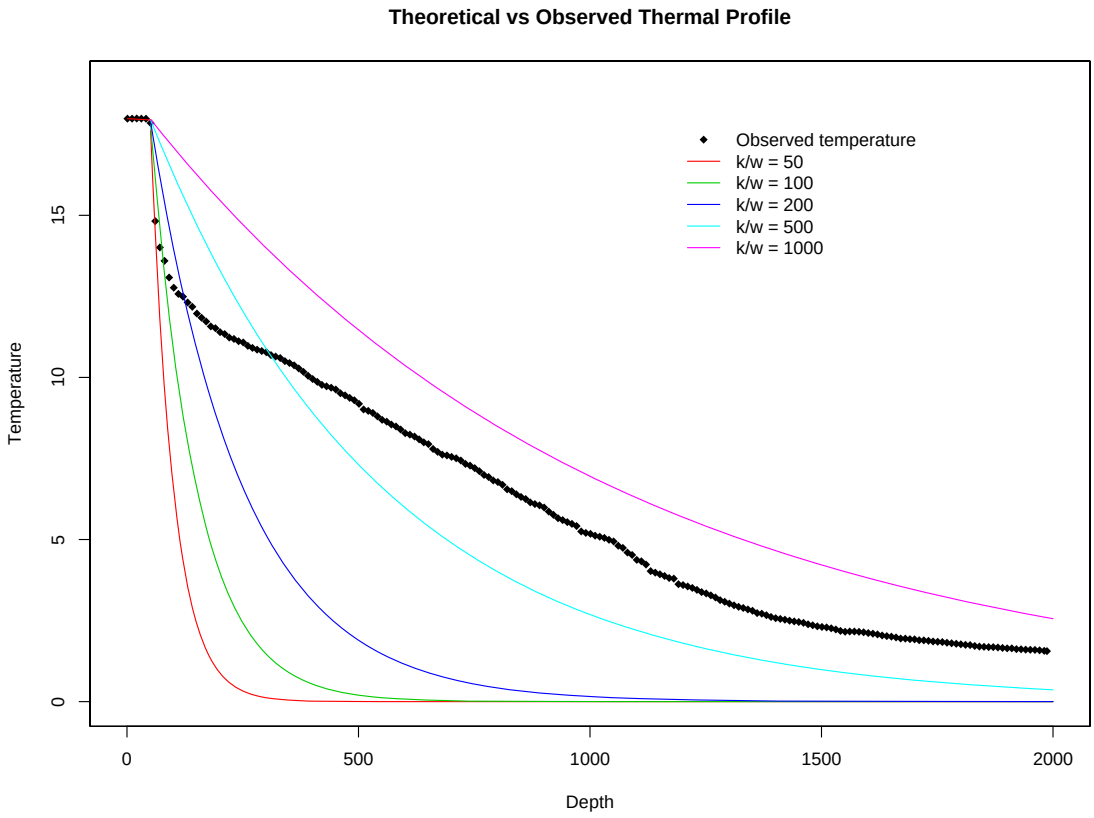


Figure 2.3: Illustration of the relationship between theoretical and observed temperature profiles for various values of k/w

most striking one is the lack of fit for the exponential model. Although one could perform statistical analyses of the observed profiles to learn constants $a, b, w,$ and k, it is our opinion that there is too much variability in the data to identify their values consistently.

It is not always possible to identify M as clearly as in figure 2.3. As an example, figure 2.4 illustrates the ambiguity about M in some observed profiles. Teal, pink,

green, and blue profiles all have more than one modal value for M. Without studying other profiles nearby, the oceanographer cannot clearly distinguish between the modes. It may be that the shallower $\hat{M}$ is due to daily warming and that the real M is closer to the deeper $\hat{M}$. Another possibility is that the shallower $\hat{M}$ is the true M while the deeper $\hat{M}$ is a remnant from the previous season. The first scenario is far more likely to hold for the green profile than for the blue profile. It is more likely that the temperature difference between the two $\hat{M}$'s is better explained by the daily warming in the green profile than in the blue profile. It could also be that the current M truly is the shallower of the two, and that the second mode is where M used to be earlier in the year, but waters have not had a chance to mix completely yet. The red profile, on the other hand, exhibits a different kind of ambiguity. In this case, cooling of the surface temperature is apparent, creating an unstable situation where cooler waters are sitting on top of warmer waters.

It is important to note that increasing the sampling resolution in any of these profiles would not necessarily resolve the ambiguity. The features of these profiles, *i.e.* the departure from the overall large scale M process, are characterized by thermodynamic processes in the ocean caused by daily warming, nightly cooling, *etc*, and not by the lack of temperature data. Whether the goal is interpolating or extrapolating, a valuable model describing the data generating process must include these factors. Mathematical modelers are working hard with little success on reproducing temperature profiles and M. From a statistical point of view, such modeling is not feasible and is the reason that ordinary parametric models will not yield the desired uncertainty. In the following chapter, we will introduce a mechanism for measuring evidence provided by the data using the expert's opinion.

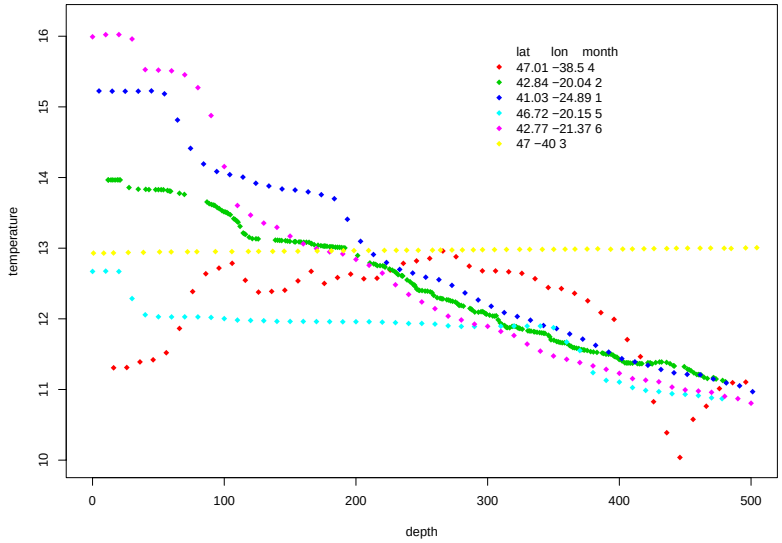


Figure 2.4: Some examples of irregular temperature profiles.

2.1.2 Temporal and Spatial dependence

Since the formation of the Mixed Layer is influenced by energy forces at the surface, it is natural to expect that M should vary seasonally. In fact, the annual cycle of the Mixed Layer depth is opposite to the annual cycle of the surface temperature. The Mixed Layer is deepest in the winter and shallowest in the summer. Figure 1.1 is a good illustration of the smooth evolution of a profile at any location as a function of surface temperature.

The annual cycle of the Mixed Layer can also be expected to vary with latitude. Near the equator, the annual cycle should be barely noticeable due to the almost constant inter-annual heat flux while the transition periods of warming and cooling should be expected to last several months. On the other hand, at higher latitudes the annual cycle should be well pronounced; M should be deeper, the transition periods should occur more rapidly, and the warm periods should be shorter. The seasonal variability and its dependence on latitude is visible from the observed data. In figure

2.6, we plotted the observed profiles at mid latitudes in the subtropical gyre, and, in figure 2.7, we plotted the observed profiles at higher latitudes. We see longer periods of deep mixing and a quicker transition to shallow mixing at higher latitudes.

The ambiguity mentioned in reference to figure 2.4 may be resolved by other profiles taken at the same or nearby location within a short period of time. For example, let us look at the profiles in figure 2.5. Based on the first observed profile marked by blue lines, the oceanographer believes $M = 0$, but she also gives some probability to the interval $(70, 80)$ and explains the slope on $(0, 70)$ as daily warming. We observe the green profile approximately 20 km north of the blue profile. For this thermocline, the oceanographer concludes $M \in (50, 60)$. Combining the information provided by both profiles updates the oceanographer's assessment about the first profile. She now believes M is at the surface in the blue profile, since it is unlikely that the waters further south(blue) would mix deeper $(70, 80)$ than the waters of the green profile. Similarly, if we were to observe the pink profile first, we could not easily distinguish between the intervals $M \in (10, 20)$ or $M \in (50, 60)$. However, after observing the green profile, we may believe $M \in (50, 60)$ is closer to the truth. While the formation of the Mixed Layer in the warming months is greatly influenced by heating at the surface, the stratification of temperature profiles in the months of cooling is driven by different processes, but it is primarily by convection of the waters. In the left panel of figure 2.5, we plotted two profiles taken approximately $50km$ apart in the month of November. During the stratification, the mixed layer is well pronounced even while the entire profiles are cooling.

In the following chapters, we will be concerned with providing a good method for weighing evidence about M provided in the data. This will enable us to model systematic dependencies in M, such as spatial proximity and the inter-annual cycle, in order to assess the evidence of long term variability.

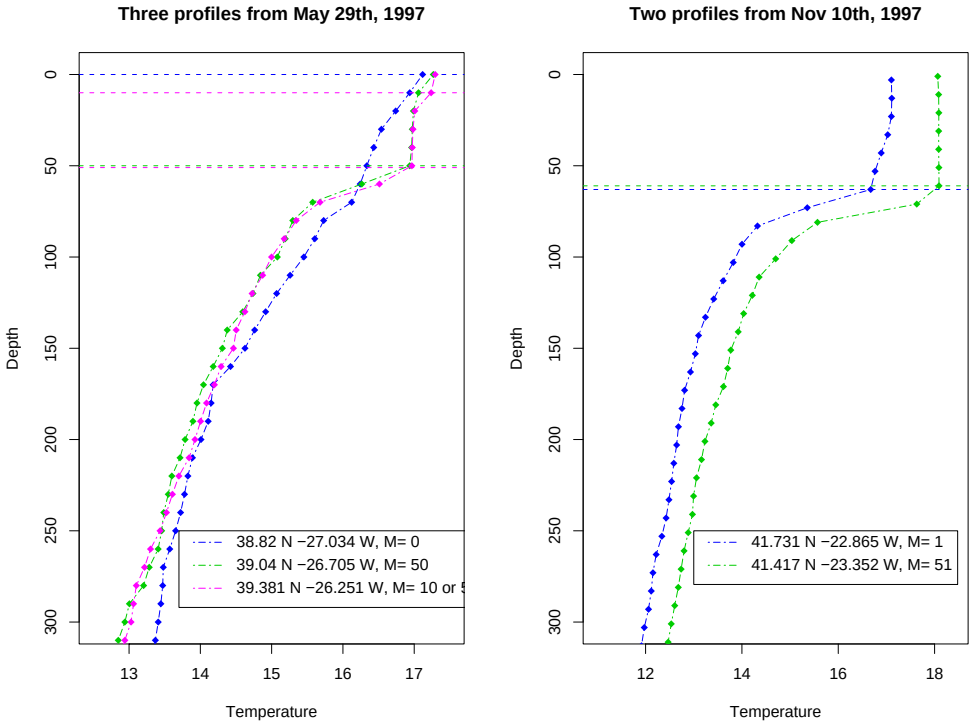


Figure 2.5: Examples of daily variability of MLD.

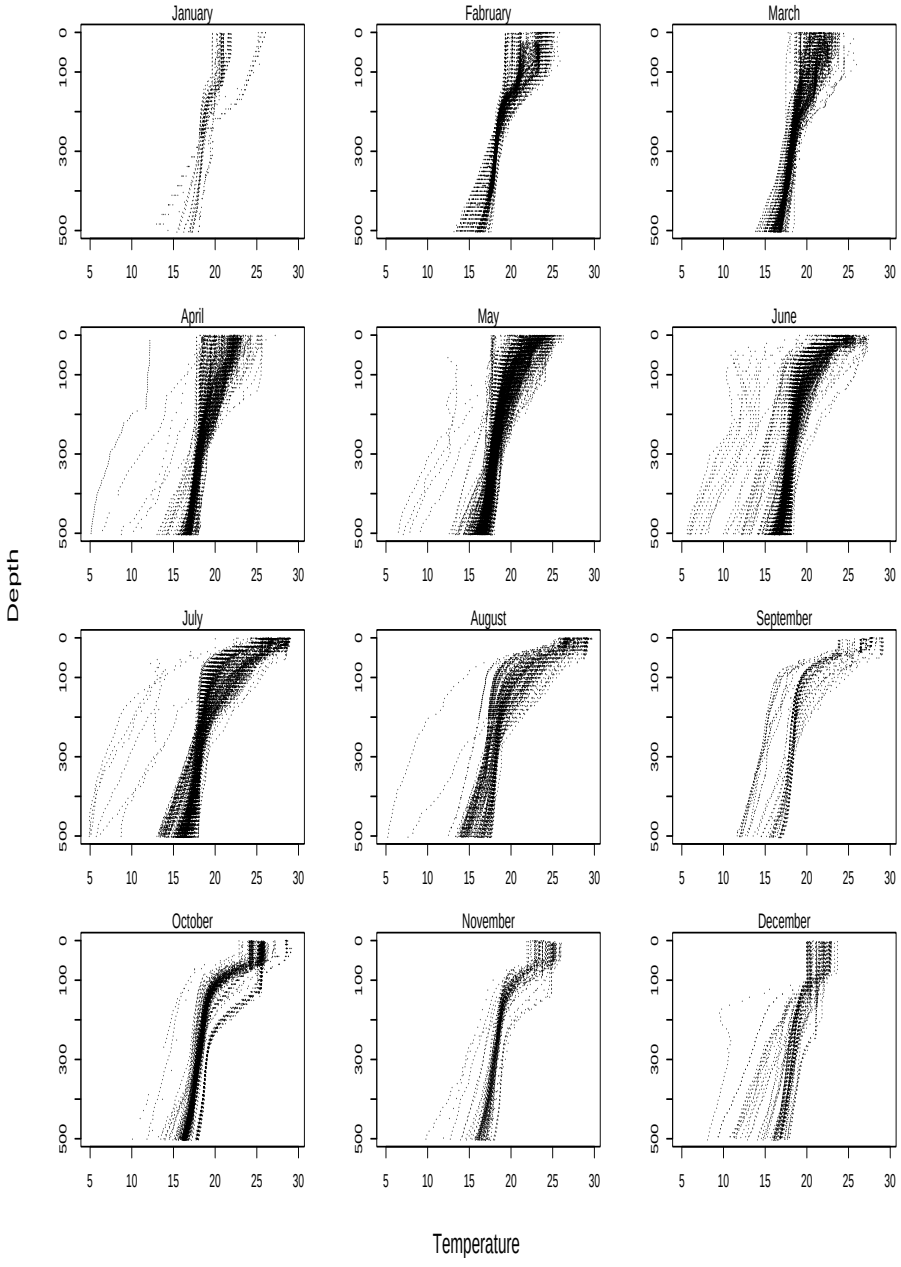


Figure 2.6: Observed thermal profiles in the subtropical gyre; profiles are grouped by months of the year: by rows and top-down.

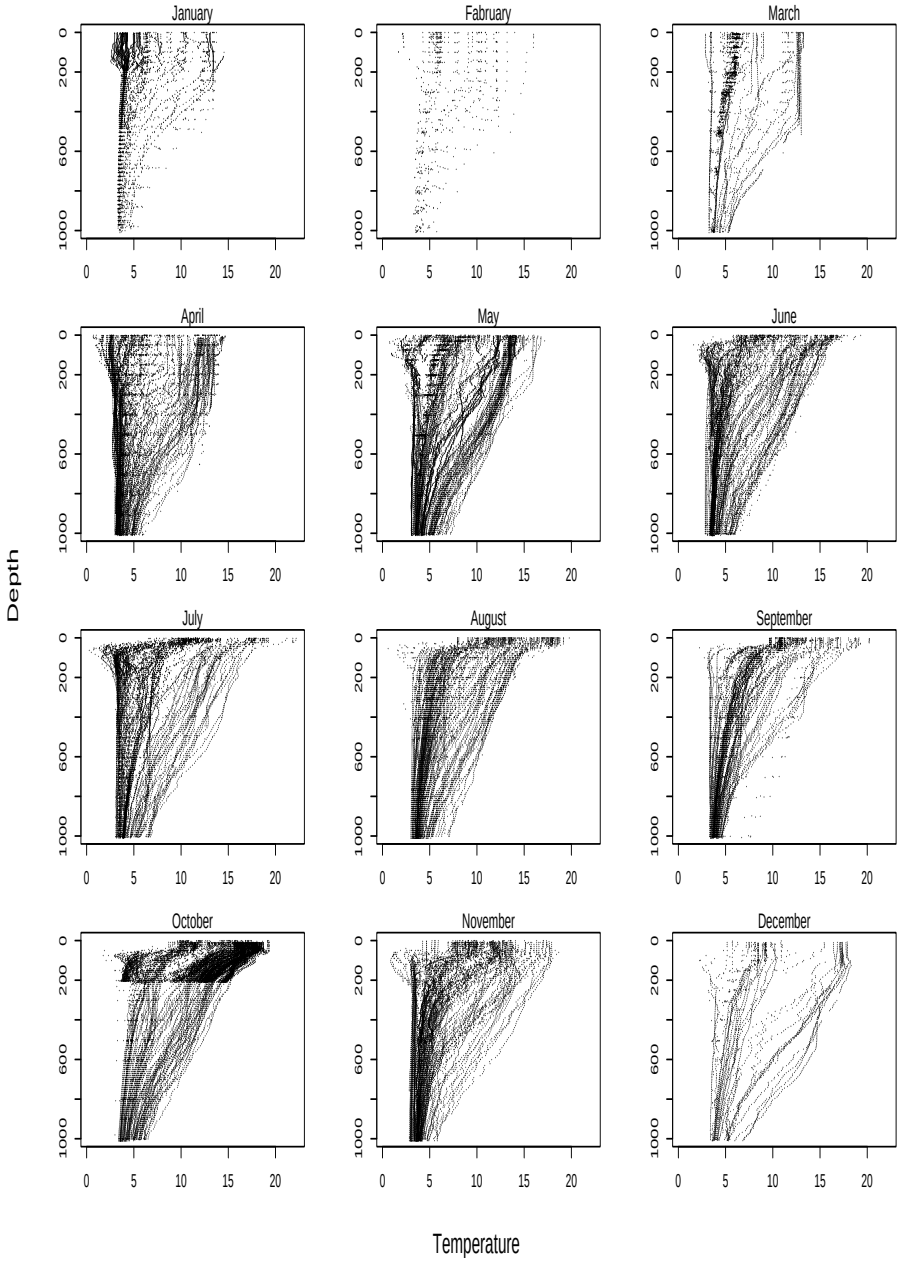


Figure 2.7: Observed thermal profiles in the subpolar gyre; profiles are grouped by months of the year: by rows and top-down.

Chapter 3

Algorithmic Likelihood

3.1 In the following order: Prior, Posterior, Likelihood

Suppose that there exist two random quantities X and θ with some joint distribution, $P(X, \theta)$, and that we are interested in the inference about the unobservable quantity, θ, given the observed realizations $X_1, X_2, X_3, \cdots$. The data generating mechanism, *i.e.* the sampling distribution, is not known and is assumed to be non-trivial; $X_1, X_2, X_3, \cdots$ may have an arbitrary dimension, they need not be independent, θ may be indexed by the same variable as X, and, moreover, we are quite certain that θ does not uniquely determine all unknown features of X. Before and after seeing any realization $X = x$, the expert can evaluate her prior and posterior belief about θ. However, it is infeasible for the expert to draw inference on θ simultaneously for all $X_1, X_2, X_3, \cdots$ without a probability model. She would like to assign a posteriori conditional distributions $\theta | X_1, X_2, X_3, \cdots$, but she doesn't have a likelihood function.

In this chapter, we propose a scheme for obtaining a likelihood-like function based on the expert's posterior distribution. The expert's posterior conditional distributions have similar interpretations to those obtained by the fiducial argument with

subtle, but important, differences. In fact, the conditional distributions adhere to the relevancy principle of Dempster (1963), and should, perhaps, be called *direct probabilities*.

The scheme is as follows: In consideration of the desired inference, create a set of partitions in θ. Considering the goals of inference and the experimental design, these partitions may depend on the observed data or some aspects of it. Present the data to the expert. Ask her to express her posterior belief about θ on a given partition based only on the data in front of her. She should not be given the kind of knowledge that distinguishes $P(X|\theta_i)$ from $P(X|\theta_j)$ for $i \neq j$. Find a functional form, $h(X, \theta)$, bounded in θ, that yields a posterior uncertainty approximately equal to that of the expert. Define a transformation of variables, $(X, \theta) \rightarrow U$, whose joint distribution is given by $h(X, \theta)$. The transformed variables are called pivotal random variables, and they imply the joint distribution $P(X, \theta)$. If the distribution of the pivotal quantities is regarded as independent of the data X, then the posterior conditional distribution, $P(\theta|X)$, is called a *direct probability* distribution. If the distribution of the pivotal variables is regarded as independent of θ, $P(\theta|X)$ is called a *fiducial distribution*

$$h(X, \theta) \quad \propto \quad -\frac{\partial}{\partial \theta} F(X, \theta) d\theta. \tag{3.1}$$

The two distributions are equal, but the latter is not a probability distribution.

A third conditional distribution, Bayes' posterior, arises if the pivotal quantities are regarded as independent of θ, and the marginal distribution $\pi(\theta)$ is specified.

The three are equal if, and only if, $f(x|\theta)$ belongs to the exponential family, and θ is a location parameter with the un-normed Lebesgue measure as a marginal distribution (Lindley, 1958).

The posterior distribution of the expert arises as the distribution of the pivotal variables after observing the data. This is no other than the fiducial distribution, or

the likelihood as a function of θ. Therefore, when the data model is not known, we propose to use the expert's posterior in place of the likelihood. Furthermore, with the substitute likelihood, we can model the apriori knowledge or other dependencies in $X_1, \cdots, X_n$ through the hierarchical specification.

The function $h(X, \theta)$ is not equal to the product of conditional densities of X, and it may not have properties such as a sufficient statistic, a normalizing constant, *etc.*

Because of the application-based nature of $h(X, \theta)$, in the following sections we introduce a particular $h(X, \theta)$ within a motivating example, and then we follow with a discussion as to why the algorithmic likelihood may be of much broader use than first imagined. In a later chapter, we will introduce a reasonable modeling approach to the problem and then follow up with a discussion comparing the results.

3.2 A Motivating Example

We observe a discrete realization $X = x = (x(d_1), x(d_2), \cdots, x(d_n))$ of a temperature profile $x(d)$ at equal distance depths $d_1, \cdots, d_n$, and we are interested in the inference about the depth of the Mixed Layer M. Depths of the measured temperatures are known and considered a part of the experimental design. The uncertainty about M is not associated with n. It is assumed that measurements are taken accurately, so all the ambiguity about M at a particular location is due to thermodynamic processes in the ocean. In the final chapter, we will model profiles taken at different locations as exchangeable, given the spatio and temporal marginal distribution of $M(s, t)$. In this section, we are concerned with establishing $P(x(d)|M)$ at a particular location in time and space.

A priori, without knowing day, year, latitude, or longitude, the oceanographer thinks M can be anywhere between 0 and B, the surface and some depth, respectively.

Shown the data, again without day, year, latitude, or longitude, the oceanographer updates her belief about M. We ask her such questions as:

a) Which observed interval(s) (d_i, d_{i+1}) is M most likely to be in?

b) How much more likely is it that M is in the interval $[d_i, d_{i+1})$, as opposed to the interval $[d_j, d_{j+1})$ for $j \neq i \in \{1, \cdots, n-1\}$?

Answers to these questions give us the expert's posterior probability, $P(M|X)$. While explaining her posterior belief, the oceanographer says her opinion is based on two pieces of information: let $i = \max\{k;\ d_k < M,\ \ k = 1 \cdots n\}$, then

I. $\triangle_1(M) = x(d_1) - x(d_i)$,
the difference in temperature between d and the surface, and

II. $\triangle_2(M) = \frac{x(d_i) - x(d_{i+1})}{d_{i+1} - d_i}$,
the rate of cooling around $d \in [d_i, d_{i+1})$.

Our approach is to fit a function, $h(M, X) = h(\ \triangle_1(M),\ \triangle_2(M)\ )$, to the oceanographer's assessments, and set it equal to the posterior distribution $c\,P(\ M \,|\, x(d)\ )$. After trying several functional forms, we settled on:

$$h(M) \propto \exp\{-\delta_1\}^{\alpha}\ \times\ [1 - \exp\{-\,\delta_2\,\beta\}]\,, \tag{3.2}$$

where

$$\delta_1 = max\{\,0, \triangle_1(M)\,\},$$

$$\delta_2 = max\{\,0, \triangle_2(M)\,\},$$

and α and β are fixed scale parameters. The oceanographer does a good job in distinguishing among the likely events so h is a good approximation on these intervals. However, she cannot consistently distinguish among the intervals that she believes to

be unlikely. Instead, she assigns a total probability to all such intervals. We choose the posterior distribution to be

$$P(M|x(d)) = max(h(M), c_p)$$

, where the constant $c_p \in (0, 1)$ is chosen such that all intervals with $h(M) \leq p$ are assigned equal mass c_p.

$$c_p = \max_{i=1,\cdots,n-1} \{\ h(M)\ :\ P(\ M\ \in\ (d_i, d_{i+1}] \mid x(d)) \leq p\}.$$

In other words, we don't differ between the intervals of the smallest probabilities. We performed a simulation study for some good choices of p, typically $\leq 5\%$.

The first term of the posterior (3.2) penalizes large deviations from the surface temperature while the second term penalizes small drops in temperature immediately below M. This rationale of the posterior is illustrated in figure 3.1. The most likely interval to which M belongs is characterized by the left end temperature $t(d_i)$ being close to the surface temperature and a large gradient between $x(d_i)$ and $t(d_{i+1})$. Parameters α and β cannot be 'learned' by the data; instead, we must fit their values using training samples. We propose to treat $h(d)$ as a likelihood function. When we show the profiles to the oceanographer, we do not tell her the location or date of the profile. Under such conditions, it is reasonable to assume that she used $\pi(M) \propto 1$ as the prior. Additionally, since the temperatures were measured at equal depth intervals, posteriors must be proportional to the likelihood function of the parameter M.

Another way to interpret $h(M)$ as a likelihood function begins with the usual formulation of the prior $\pi(M)$ and an unknown sampling distribution

$$P(x(d_1), s(d_2), \cdots, x(d_n)|M).$$

Though P is unknown, there is, in principle, a joint distribution

$$P(x(d_1), x(d_2), \cdots, x(d_n), M).$$

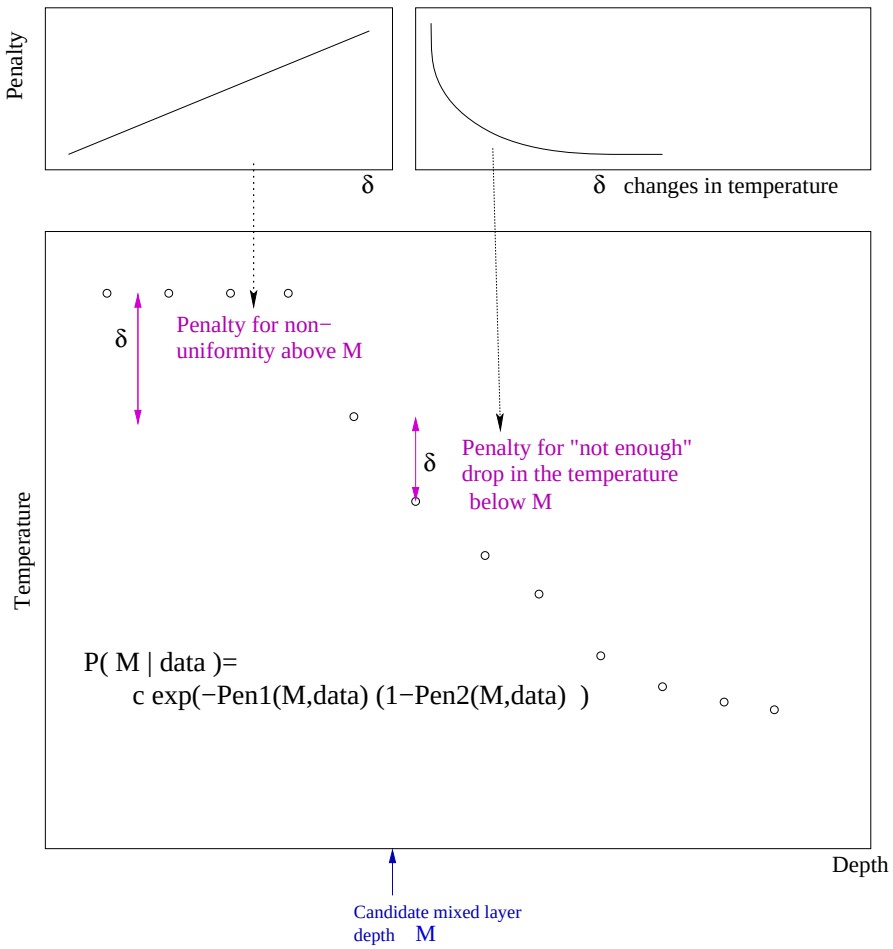


Figure 3.1: An illustration of $\triangle_1(d), \triangle_2(d)$.

We choose the $n+1$ dimensional transformation of variables

$$(M, t) \rightarrow (\triangle_1(M), \triangle_2(M), \mathbf{s}(M), M)$$

, such that the joint distribution of $\mathbf{s}(M)$ is approximately independent of the distribution of M.

Consider a theoretical relationship, described in equations 2.1 and 2.2, to be the underlying process and $x(d)$ a discrete realization of such a process observed at equal

distance depths $d_1, \cdots, d_n$. Define $i(M) = \max\{k;\ d_k < M,\ \ k = 1 \cdots n\}$ and let

$$s(M)_j = x(d_j)/x(d_{i(M)+1}) \qquad j < i(M),$$

$$s(M)_j = \frac{x(d_{j+1})}{x(d_{j+2})} \qquad j \geq i(M) \qquad j = i(M), \cdots, n-2,$$

$$\triangle_1(M) = x(d_1) - x(d_{i(M)}) \qquad s.t. \qquad M \in [d_{i(M)}, d_{i(M)+1}), \quad \text{and}$$

$$\triangle_2(M) = \frac{x(d_{i(M)}) - x(d_{i(M)+1})}{d_{i(M)+1} - d_{i(M)}} \qquad s.t \qquad M \in [d_{i(M)}, d_{i(M)+1}).$$

Since within the mixed layer $x(d) \approx x(0)$, and therefore $x(d) \approx x(M)$, it follows that for $j < i(M)$

$$s(M)_j \approx \frac{\exp\{-w(M-M)/k_1\}}{\exp\{-w(d_{i+1} - M)/k_2\}}.$$

Hence, in the upper layer of the ocean, the transformation $s(M)_j$ depends on the distance from M and not on the value of M, per se. For $j \geq i(M)$, the transformation $s(M)_j$ is also independent of the distribution of M,

$$s(M)_j = \frac{\exp\{-w(d_{j+2} - M)/k_2\}}{\exp\{-w(d_{i+1} - M)/k_2\}} = \exp\{-w(d_{j+2} - d_{i+1})/k_2\},$$

and if $d_i's$ are observed at equal distance intervals, $s(M)_j$ are identical for all j.

The joint distribution of $\triangle_1(M)$ and $\triangle_2(M)$ is independent of $x(d)$ and is expressed by $h(\triangle_1(M), \triangle_2(M))$. The transformation $\mathbf{s}(M)$ depends on M but not on its distribution.

Since the joint distribution of $P(x(d_1), x(d_2), \cdots, x(d_n), M)$ on $\mathcal{X} \times \mathcal{M}$ exists, so does the joint distribution $\pi(\triangle_1(M), \triangle_2(M), \boldsymbol{s}(M), M)$ on $\mathcal{U} \times \mathcal{S} \times \mathcal{M}$. The two are

proportional up to a Jacobian constant of transformation and may be factored as

$$\begin{aligned}\pi(\triangle_1(M), \triangle_2(M), \mathbf{s}(M), M) &= \pi(\triangle_1(M), \triangle_2(M)|M) \\ &\quad \times \pi(\mathbf{s}(M) \,|\, \triangle_1(M), \triangle_2(M), M\,)\pi(M).\end{aligned}$$

However, since $\triangle_1(M)$ and $\triangle_2(M)$ contain little information about $\mathbf{s}$, we say:

$$\pi(\triangle_1(M), \triangle_2(M), \mathbf{s}, M) = \pi(\triangle_1(M), \triangle_2(M)|M)\,\pi(\mathbf{s}|M)\pi(M).$$

After observing the data $x(d_1), x(d_2), \cdots, x(d_n)$, the conditional distribution of M is inferred from the transformed variables $\triangle_1(M)$ and $\triangle_2(M)$ rather than the original observed values in $\mathcal{X} \times \mathcal{M}$:

$$\begin{aligned}P(M|\, x(d_1), x(d_2), \cdots, x(d_n)) &\propto \pi(\triangle_1(M), \triangle_2(M)|M)\pi(\mathbf{s}|M)\pi(M) \\ &= h(\triangle_1(M), \triangle_2(M)|M)\;\pi(M)\end{aligned}$$

Regardless of whether such a prior $\pi(M)$ is known (or exists), $h(\triangle_1(M), \triangle_2(M))$ represents the conditional probability of the data.

The transformation $(M, x(d)) \rightarrow (g(M, x(d)), \mathbf{s})$ is designed to describe the oceanographer's belief that $\mathbf{s}$ contains little information about M. Because the observed temperatures do not exactly decay at an exponential rate below the Mixed Layer, the observed values of s_j, where $j \geq i(M)$ and $i(M) = \max\{\, k\,;\; d_k < M,\;\; k = 1 \cdots n\,\}$, may not be identical. Therefore, one can argue that $\mathbf{s}$ contains some information about M and cannot be regarded as independent from $\triangle_1(M)$ and $\triangle_2(M)$. A transformation on the theoretical profile from Chapter 2 is given in figure 3.2, and transformations on the three observed profiles are given in figure 3.3. To reassure the reader that this loss of information is minimal, we simulated several scenarios:

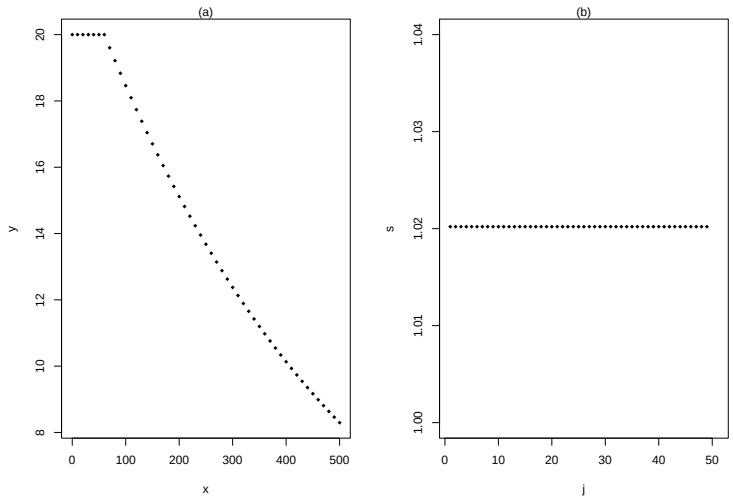


Figure 3.2: An illustration of transformation **s** of a theoretical profile in the case where discretization comes in equal length intervals.

1. Let observed depths $d = (0, 10, 20, 30, 40, 50)$, $i = 3, \alpha = 1.0, \beta = 1/0.3$, $\triangle_1(M) = 0.01$, and $\triangle_2(M) = .15$. This translates to a $0.01\,^oC$ change between the surface and the bottom of the Mixed Layer, and $1.5\,^oC$ decay between $[20m, 30m)$. We then simulated 100 samples of **s** from an arbitrarily chosen $\mathcal{U}(1, 3)$ such that $s_1 = s_2 = s_3 = s_4$, as is assumed by the theoretical relationship. The resulting profiles and their likelihood functions are plotted in the top row of figure 3.4. The profiles are normalized to fit a temperature range between 0 and 1 since $s \in (1, 3)$ accounts for a very large range of temperatures.

2. Let observed depths $d = (0, 10, 20, 30, 40, 50), i = 3, \alpha = 1.0, \beta = 1/0.3, \triangle_1(M) = 0.01$, and $\triangle_2(M) = .15$. We simulate s_2, s_3, s_4 from $\mathcal{U}(1, 3)$ as three distinct *iid* random variables, and set $s_1 \equiv s_2$. This scenario corresponds to the observed transformations. The resulting profiles and their likelihood functions are plotted in the bottom row of figure 3.4.

3. Let d, α, β, and **s** be as in the first scenario, but change $\triangle_1(M)$ and $\triangle_2(M)$ to

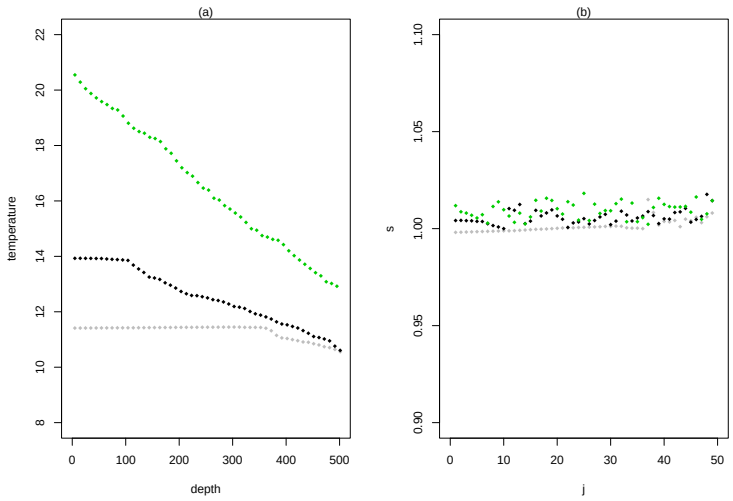


Figure 3.3: An illustration of transformation **s** on observed profiles in the case where discretization comes in equal length intervals.

0.2 and 0.05, respectively. Values of **s**, relative to those of $\triangle_1(M)$ and $\triangle_2(M)$, are such that the transformation back to temperatures will give profiles with very ambiguous evidence about M. The resulting profiles and their likelihood functions are plotted in the top row of figure 3.5.

4. Let d, α, β, and **s** be as in the second scenario, but let $\triangle_1(M) = 0.2$ and $\triangle_2(M) = 0.05$. Just like in scenario 3, the values of **s** will give profiles containing weak evidence about M. The resulting profiles and their likelihood functions are plotted in the bottom row of figure 3.5.

5. Finally, take $n = 6$, $i = 3$, fix **s** conditionally on M, and sample $\triangle_1(M)$ and$\triangle_2(M)$. The resulting rescaled profiles, respective likelihood functions, and the joint distribution of $\triangle_1(M)$ and $\triangle_2(M)$ are plotted in figure 3.6.

We created profiles in the first two scenarios with realistic values for $\triangle_1(M)$ and $\triangle_2(M)$, and the depth of the Mixed Layer is well pronounced. There is very little

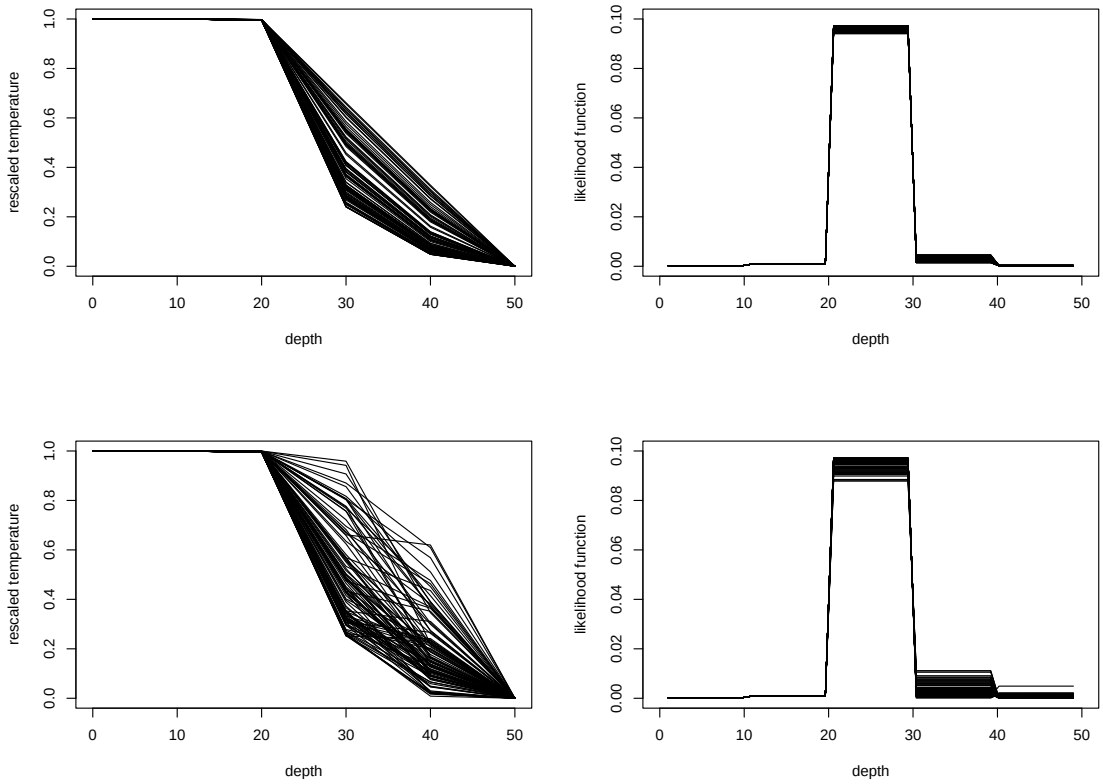


Figure 3.4: Simulated profiles for fixed values of $\triangle_1(M) = 0.01$ and $\triangle_2(M) = .15$ and their respective likelihood functions. Top row: transformation $\mathbf{s}$ as in the theoretical relationship, bottom row: $\mathbf{s}$ as typically observed.

difference between the theoretical relationship (top row) and the observed transformation (bottom row), indicating little loss of information by assuming $s_1 = s_2 = s_3 = s_4$. In scenarios 3 and 4, the values of $\triangle_1(M) = 0.2$ and $\triangle_2(M) = 0.05$ indicate a less pronounced Mixed Layer, and there is more variability in the likelihood function than in scenarios 1 and 2. More specifically, values of $\mathbf{s}$ are simulated from such a wide range, relative to $\triangle_1(M)$ and $\triangle_2(M)$, that the resulting profiles will appear to have a Mixed Layer in the intervals other than $[20m, 30m)$. Although there is more variability in scenario 4 (observed), relative to that of scenario 3 (theoretical), the likelihood function in the top row remains robust. The increase of variability

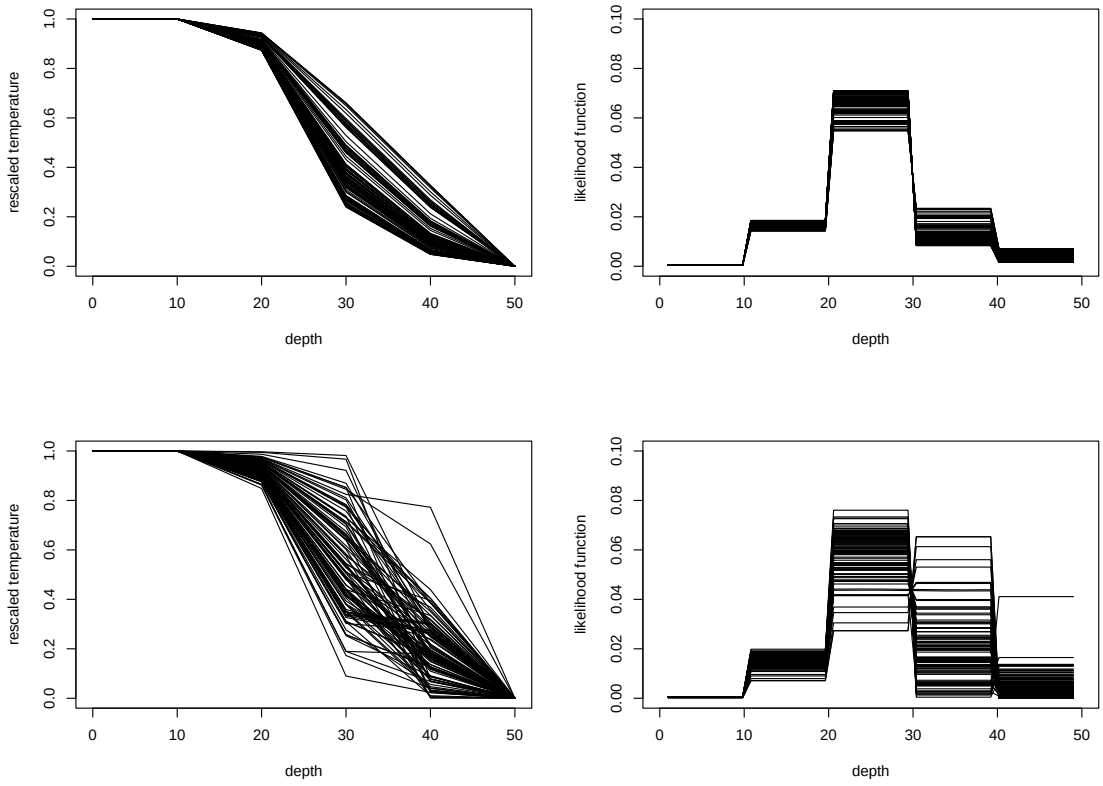


Figure 3.5: Simulated profiles for fixed values of $\triangle_1(M) = 0.2$ and $\triangle_2(M) = 0.05$ and their respective likelihood functions. Top row: transformation $\mathbf{s}$ as in the theoretical relationship, bottom row: $\mathbf{s}$ as typically observed

among likelihoods in 3.5, relative to those in 3.4, points out that the likelihood is more sensitive to the choice of $\triangle_1(M)$ and $\triangle_2(M)$ rather than the transformation $\mathbf{s}$. Finally, in the last scenario, we fix the value of $\mathbf{s} = s_1 = s_2 = s_3 = s_4$ and i, and simulate $\triangle_1(M)$ and $\triangle_2(M)$ according to $h(M)$. The likelihood functions in 3.6 should, now, be compared with those in the top rows of 3.4 and 3.5. We see that, under the assumption $\mathbf{s} = s_1 = s_2 = s_3 = s_4$, the likelihood function reflects the uncertainty due to the choice of $\triangle_1(M)$ and $\triangle_2(M)$.

There can be yet another way to interpret our likelihood: let M be the parameter of interest, and let θ_M be the parameters that fully specify the distribution $(x|\theta_M, M)$

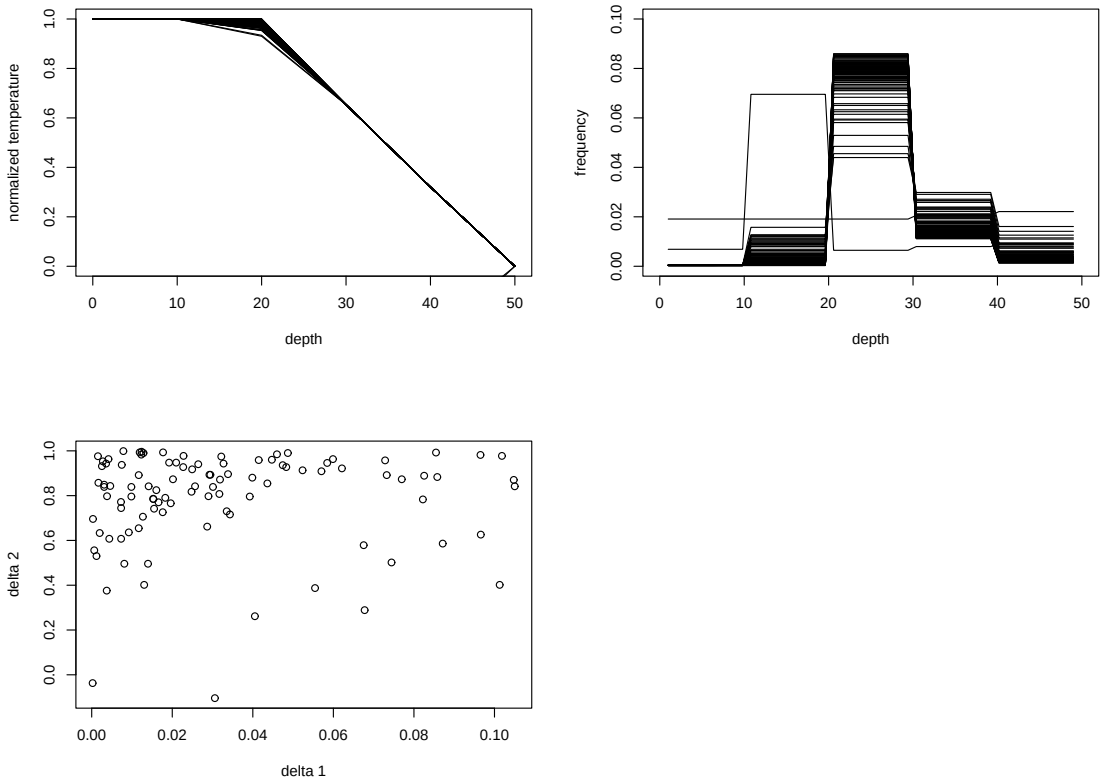


Figure 3.6: Simulated profiles for fixed values of **s**, their respective likelihood function and sampling distribution of $\triangle_1(M)$ and $\triangle_2(M)$

for any M. Then

$$\begin{aligned} P(x_1, \cdots, x_n|M) &= h(\triangle_1(M), \triangle_2(M)) \\ &= \pi(M) \times \int (x_1, \cdots, x_n|\theta_M, M) P(\theta_M|M) d\theta_M. \end{aligned}$$

The likelihood function may, in this case, be interpreted as a marginal probability of the data given the model M, and the elicited probabilities are closely related to Bayes factors.

3.3 Some related ideas

The idea of equating likelihood and the posterior density appeared in the works of Lindley (1958), Dempster (1963), Williams (1963), and others. These discussions are generally motivated by finding the correspondence, or lack thereof, between Fisher's fiducial distribution and the Bayesian posterior. In a single dimension case, Fisher defines the fiducial distribution (Fisher, 1956, pg. 70) as

$$\phi(\theta) = -\frac{\partial F}{\partial \theta} d\theta.$$

Associated with $X_1, \cdots, X_n$ and θ, there exist n pivotal variables whose distribution is known and independent of θ. That is because there exists a transformation of pivotal variables such that only one of the transformed sets $P(T, \theta)$ involves θ, and the others do not (Barnard, 1963). Transformation T is said to be sufficient for θ. Then, if (Fisher, 1956; Brillinger, 1962)

a. $F(T, \theta)$ is a distribution function of a sufficient statistic T,

b. θ and T vary continuously over the same range,

c. F is a strictly monotonic function of θ for any fixed value of T, and

d. F is a strictly monotonic function of T for any fixed value of θ,

the probability space on θ arises from the distribution of T after observing $x_1, \cdots, x_n$. The distribution of T remains independent of θ even after the data is observed

Fiducial inference was introduced as a concept and was not supported with rigorous theory. It is intended for the post-data interpretation of evidence and clearly provides a probabilistic interpretation for θ. Apparently, Fisher himself had probabilistic interpretations in mind:

> ... The concept of probability involved [in the fiducial argument] is entirely identical with the classical probability of the earlier writers, such as Bayes. *Fisher(1956) pg. 51.*

However, without accepting a probability model for the involved parameters, it is difficult to justify any probabilistic statements, but the fiducial argument clearly had an attractive interpretation (post data), so for some time statisticians were concerned with the conditions under which a fiducial argument was true. Grundy (1956) gave examples of one-parameter distributions for which a fiducial distribution does not equal the posterior for any choice of prior. Grundy's argument is based on the sampling distribution of x:

$$\frac{1}{\sqrt{(2\pi)}}\exp(-\frac{1}{2}(\theta - x)^2)\mathbf{f}(x)/g(\theta),$$

where $\mathbf{f}(x) \geq 0$ is a polynomial of order greater than 0, and $g(\theta)$ is an un-normed Lebesgue measure. Other inconsistencies were found, such as non-uniqueness due to the choice of a pivotal variable, two dimensional problems, etc. In fact, Lindley (1958) gives a necessary and sufficient condition for the fiducial distribution to be equal to the posterior distribution. The condition requires that there is a transformation of data from x to u and of parameter θ to τ, where τ is a location parameter for u. Furthermore, if a single sufficient statistic exists for θ, it must satisfy that condition. As a result, only five single parameter exponential family distributions satisfy this condition; normal mean, single parameter gamma, negative binomial, Poisson, and binomial (Kendall *et al.*(1991, sec 31.39)). These are also the families where a transformation invariant measure on θ exists, the so called objective priors. Fiducial inference lost its popularity, but many statisticians continue to appreciate the concept. Barnard (1963) explains the fiducial distribution using group theory arguments, and a part of the Bayesian community (those interested in the invariant

priors) still makes reference to the work. Dempster (1963) gives a different view in his argument for *Direct Probabilities*. His idea is that the probability space on θ arises naturally through the distribution of pivotal quantities if their distribution is not regarded as independent of θ as assumed in a fiducial argument but independent of X instead.

Although Dempster did not have in mind substituting the likelihood with a posterior distribution, as in our case, he did have in mind inferring the posterior directly from the distribution of the pivotal quantities. He gave a theorem explaining why this is possible. Therefore, we think it is worth describing Dempster's relevancy principle here.

The argument starts with accepting a joint probability model for the observed and unobserved random variables through an associated set of pivotal variables and their known distributions. After observing X, the unobserved variables are assigned an *a posteriori* distribution according to the *relevancy principle* from the distribution of the pivotal variables. These probabilities are called *Direct Probabilities*. The relevancy principle asserts the use of inverse probabilities, but it does not require that the *a priori* distribution to be a function of only the unknown parameters. Therefore, it inherently differs from the Bayesian posterior distribution. It also differs from the frequentist interpretation, such as E.S. Pearson and J. Neyman, because it adopts a global probability model. In fact, *Direct Probabilities* are most closely associated with Fisher's fiducial distributions and may be viewed as a generalization of Bayesian posteriors. An explanation comes from the concept of sampling.

Dempster (1963):

> Consider a family $\mathcal{F}_\theta$ of continuous cumulative distribution functions denoted $F(x, \theta)$.
>
> *Definition.* Random variables $X_1, X_2, \cdots, X_n$ will be called a random

sample from one of the family $\mathcal{F}_\theta$, or, briefly, a sample from $\mathcal{F}_\theta$, provided that:

i. $X_1, X_2, \cdots, X_n, \theta$ and, hence, any functions thereof are jointly distributed random variables, and

ii. the random variables $U_1 = F(X_1, \theta), \cdots, U_n = F(X_n, \theta)$ are marginally distributed like n independent $U(0,1)$ random variables.

For families $\mathcal{F}_\theta$ where θ takes values in a real continuum, the following theorem will clarify the general method of construction of a random sample.

Theorem 3.1. Suppose $X_1, X_2, \cdots, X_n$ constitute a sample from $\mathcal{F}_\theta$. Suppose further that the conditional distribution of θ given $U_1, U_2, \cdots, U_n$ is continuous. Then there exists a function $g(X_1, X_2, \cdots, X_n, \theta)$ which (a) is continuously distributed independent of $U_1, U_2, \cdots, U_n$ and which (b), together with $U_1, U_2, \cdots, U_n$, determines $X_1, X_2, \cdots, X_n$ and θ completely. Conversely, given any function $g(X_1, X_2, \cdots, X_n, \theta)$ with property (b), a random sample $X_1, X_2, \cdots, X_n$ may be chosen by choosing g according to an arbitrary distribution and independently choosing $U_1, U_2, \cdots, U_n$ to be independent $U(0,1)$.

The posterior inference may depend on the value of $g(X_1, X_2, \cdots, X_n, \theta)$ but not its distribution. Moreover, as the sample size n grows to infinity, the dependence becomes increasingly less relevant, just like the importance of the prior in Bayesian inference and the lack thereof in frequentist inference.

Dempster gives two examples and is not interested in further demonstration or exploring the reasons to use the relevancy principle. Although the second example is more relevant to the work in this thesis and will be examined more closely later,

the first example illustrates these subtle points clearly. Suppose we wish to make inference on θ, where X is drawn from one of the family of populations with cdf $F(x|\theta) = 1 - e^{-\theta x}$. The associated pivotal variable U,

$$U = F(x|\theta),$$

as well as $1 - U = e^{-\theta x}$, is distributed as *iid* $\mathcal{U}(0,1)$. The joint distribution of $U, X,$ and θ is completed by requiring the distribution of U to be independent of the distribution of X. Now, let $g(X_1, X_2, \cdots, X_n, \theta) = X_1$ be drawn from some random process and U_1 from $\mathcal{U}(0,1)$. Together, X_1 and U_1 uniquely specify θ, and the rest of the sample $X_2, \cdots, X_n$ may be obtained conditionally on θ by drawing $U_2, \cdots, U_n$ independently from $\mathcal{U}(0,1)$. Then, to make inference on θ from the observed space $(\mathcal{X} \times \theta)$, we begin with the joint density of $U_1, U_2, \cdots, U_n$ and $g(X_1, X_2, \cdots, X_n, \theta) = X_1$:

$$\begin{aligned} P(U_1, U_2, \cdots, U_n, g) &\propto P(g(X_1)), \\ &\text{then} \\ \pi(X_1, X_2, \cdots, X_n, \theta) &= P(g) \times \left| \frac{\partial(U_1, U_2, \cdots, U_n, g)}{\partial(X_1, X_2, \cdots, X_n, \theta)} \right|. \end{aligned}$$

In particular, for the above example:

$$\begin{aligned} \pi(X_1, X_2, \cdots, X_n, \theta) &= P(g(X_1)) \times \begin{vmatrix} -\theta \exp\{-\theta\ x_1\} & 0 & \cdots & 1 \\ 0 & -\theta \exp\{-\theta\ x_2\} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ -x_1 \exp\{-\theta\ x_1\} & -x_2 \exp\{-\theta\ x_2\} & \cdots & 0 \end{vmatrix} \\ &= P(g) x_1 \exp\{-\theta x_1\} \theta^{n-1} \exp\{-\theta \sum_{i=2}^{n} x_i\} \end{aligned}$$

which, if $g(X_1)$ is independent of θ, yields $\Gamma(n, \sum_{i=1}^{n} x_i)$ as a conditional posterior on θ. Hence, the conditional posterior may depend on the value of g but not on its distribution since g was chosen to be independent of θ.

I. Consider a choice of $P(g(X_1)) \propto 1$, $n = 1$ and a joint model $\pi(X, \theta) = X \exp\{-\theta X\}$, as before. The posterior distribution, given by the relevancy principle, is $\Gamma(1, x_1)$. Integrating X out of the joint model yields the marginal distribution for θ:

$$m(\theta) \propto \Gamma(2)/\theta^2.$$

It is possible to find a model for the data and its density function such that, together with $m(\theta)$, we yield the same posterior:

$$f(x|\theta) = \frac{\pi(\theta, x)}{m(\theta)} \times P(g(x_1)) = x_1 \exp\{-\theta\ x_1\},$$

an exponential density.

II. Consider a different choice of $P(g(X_1)) \propto \lambda \exp\{-\lambda\ X\}$, $n = 1$, and a joint model $\pi(X, \theta) = P(g(X_1))\, X \exp\{-\theta X\}$, as before. By the relevancy principle, a posterior conditional distribution of θ is the same as in I. Integrating over X yields a different marginal distribution for θ than in I.:

$$m(\theta) \propto \frac{\lambda}{(\lambda + \theta)^2}\Gamma(2).$$

Consequently, a different model for the data

$$f(x|\theta) = x \exp\{-(\lambda + \theta)x\},$$

together with the new $m(\theta)$, will yield the same conditional posterior inference as in I.

It is now clear how the joint model on pair (X, θ), arising from the distribution of the pivotal quantities, need not imply the use of a specific prior. In fact, there need not exist a known prior distribution and data model that will produce corresponding direct probabilities (next section). Incidentally, it also brings to light that a Bayesian

specification of the model need not be unique. Dempster makes a point that it is not necessary to assume that the distribution of the pivotal variable remains independent of θ even after observing X, as in the fiducial argument. Additionally, there is no reason to assume g is only a function of θ, as in Bayesian inference. *Direct probabilities* may depend on the choice of g, but they are not dependent on it's distribution if g is not a function of θ. The exception is a location family where the choice of g is irrelevant. In later papers, Dempster (1963), Dempster (1966), and Dempster (1968), the author furthers his ideas and establishes an original thesis on probabilistic reasoning in statistics which he calls Generalized Bayesian inference. The Generalized Bayesian inference, as a special case, includes Bayesian and fiducial arguments and is an effort to bring the two arguments under the same umbrella. Without choosing a likelihood and a prior, the relevancy principle makes the posterior inference $P(\theta|X)$ possible directly from the distribution of the pivotal variables in a transformed space $(\mathcal{U}, g)$.

The second example of Dempster closely resembles our situation and warrants further discussion in the following section.

3.4 Relationship between $h(M)$ and Dempster's *Direct Probabilities*

The following paragraph is paraphrased from Dempster (1963): Suppose we are interested in sampling from a family $\mathcal{F}_\theta$ of all continuous distribution functions $F(x, \theta)$ with θ in the role of an indexing parameter. Let $X_1, \cdots, X_n$ constitute a sample from such a family of distributions, and let $U_1, \cdots, U_n$ be a marginally independent set of pivotal variables. Then, if

$$U_i = F(X_i, \theta) \text{ for } i = 1, \cdots, n,$$

by the monotonicity of F for a non-decreasing ordered sample $X_{(1)}, \cdots, X_{(n)}$,

$$U_{(i)} = F(X_{(i)}, \theta) \text{ for } i = 1, \cdots, n \tag{3.3}$$

holds. Even after conditioning on the entire sample, $g = X_1, \cdots, X_n$, the joint density of $U_{(1)}, \cdots, U_{(n)}$, $n! f_U(u_{(1)}) \cdots f_U(u_{(n)})$ remains independent of $X_1, \cdots, X_n$. Moreover, $U_{(1)}, \cdots, U_{(n)}$, together with g, does not uniquely determine the distribution of θ (condition (b) of Theorem 3.1). Instead, they determine a distribution across the sub-families in $\mathcal{F}_\theta$, which satisfies 3.3. Therefore, the relevancy principle may be applied across the sub-families of $\mathcal{F}_\theta$.

Similarly, we may define the depth of the Mixed Layer M to be in a deterministic relationship with the interval $[d_M, d_{M+1})$, where $d_M = \max\{\, i\,;\ d_i < M,\ \ j = 1 \cdots n\,\}$. Then the transformation $\mathbf{s}$, defined as before, is a function of M, but it is independent of the distribution of M, and, therefore, independent of the distribution of $x(d_M)$. We define two pivotal variables, jointly distributed according to $h(M)$, as

$$\begin{aligned} \triangle_1(M) &= x(d_1) - x(d_M) \\ \triangle_2(M) &= \frac{x(d_M) - x(d_{M+1})}{d_{M+1} - d_M} \end{aligned}$$

Transformation $\mathbf{s}(M)$ plays the role of $g(X, \theta)$, and together with $\triangle_1(M)$ and $\triangle_2(M)$, it does not uniquely determine x_M (condition (b) of Theorem 3.1). But it does specify the probability distribution across the sub-families of curves from which $x(d_1), \cdots, x(d_n)$ was drawn. The sub-families are defined by $M \in \{d_1, \cdots, d_{n-1}\}$ and within the $n-1$ member sub-family of curves satisfying a particular choice of $\triangle_1(M), \triangle_2(M)$ and $\mathbf{s}(M)$, all curves are assigned equal probabilities. Direct probabilities are given across the sub-families according to the distribution of $\triangle_1(M)$ and

$\triangle_2(M)$. Therefore,

$$\begin{aligned} P(d_j \leq M < d_{j+1}|\ x(d)\ ) &= P\{\triangle_1(M) = x(d_1) - x(d_j)\} \\ &\times \quad P\{\triangle_2(M) = \frac{x(d_j) - x(d_{j+1})}{d_{j+1} - d_j}\} \times P\{\mathbf{s}\} \\ &= h(\triangle_1(d_j), \triangle_2(d_j)) \qquad j = 1, \cdots, n-1 \end{aligned}$$

We define M to be a continuous variable; in which case, the sub-families are defined by the intervals rather than the rational points $d_1, \cdots, d_{n-1}$. The argument remains the same because all intervals have equal measure.

3.4.1 Can we use direct probabilities as an approximate likelihood?

Consider, again, Dempster's second example. We cannot assign probabilities within the subfamily of $\mathcal{F}_\theta$ satisfying equation 3.3 since we cannot even determine the cardinality of the set. However, we can assign direct probabilities to some important functionals such as quantiles (Dempster, 1963): let $x_p(\theta)$ denote the p^{th} quantile of interest

$$\begin{aligned} P(X_{(i_1)} < x_p(\theta) < X_{(i_2)}) &= P\{F(X_{(i_1)}, \theta) < p < F(X_{(i_2)}, \theta)\} \\ &= P(U_{(i_1)} < p < U_{(i_2)}) \\ &= I_p(i_1, n - i_1 + 1) - I_p(i_2, n - i_2 + 1) \\ &= P(Y < i_2) - P(Y < i_1) \ \text{ where } \ Y \sim Bin(p = 1/2, n) \\ &= P(Y = i_1) \end{aligned} \tag{3.4}$$

where

$$I_p(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^p t^{a-1}(1-t)^{b-1} dt.$$

These probabilities correspond exactly to those obtained by the fiducial argument, but so far, it is not known if any Bayesian model would produce the same distribution (Hill, 1968). Therefore, if a statistician, favoring the Bayes principle, wishes to make inference on the median given the sample drawn from an unknown distribution, he is at odds.

Jeffreys (1961, sec. 4.4) proposed using the approximate likelihood:

$$lik(x_p(\theta); data) = \left(\begin{array}{c} n \\ n\, F_n(x_p(\theta)) \end{array} \right) \frac{1}{2}^n \tag{3.5}$$

where F_n denotes the empirical distribution of the sample and $x_p(\theta)$ the median. The extension to other quantiles is straightforward.

Jeffreys' approximate likelihood is not integrable in θ and is not proportional to the product of any densities, but Lavine (1995) demonstrates that such a likelihood will yield a conservative inference. He conceptualizes the conservative inference as inference based on an approximate likelihood; this distinguishes evidence provided by the data about the parameters of interest in a "conservative sense". His proof relies on the proper probabilistic construction of an approximate likelihood, namely, the binomial distribution. The proof is based on the asymptotic property that as $n \to \infty$, regardless of the form for $F(x)$, we can expect $\frac{1}{n} \sum_{i=1}^{\infty} \mathbf{1}_{x_i > x_p(\theta)} \to \frac{1}{2}$.

It turns out that this likelihood is equivalent to treating Dempster's direct probabilities as a function of θ (Abramowitz and Stegun, 1964, sec 6.7).

Now consider a more hands-on statistician faced with the same problem. He asks an expert, "How do you judge the median based on this sample?". The expert replies, "For any value $x_p(\theta)$, I consider how many observations are smaller than $x_p(\theta)$ and how many are larger." If $x_p(\theta)$ is the median, the probability of observing any single observation greater than $x_p(\theta)$ should be 1/2. Therefore, the statistician concludes, using the natural partition $x_{(1)}, \cdots, x_{(n)}$ of $x_p(\theta)$, that the expert's probability

$P(X_{(i_1)} < x_p(\theta) < X_{(i_2)})$ is the probability $P(Y = i_1)$ where $Y \sim Bin(p = 1/2, n)$. Hence, the pivotal variable the statistician is looking for is Y with a density function

$$lik(x_p(\theta); data) = \left(\begin{array}{c} n \\ i \end{array} \right) \frac{1}{2}^n$$

where $i = \sum_{i=1}^{n} 1_{(-\infty, x_p(\theta)]}(x_i)$. His posterior distribution is identical to that of Dempster, and the corresponding approximate likelihood yields conservative inference.

3.5 Discussion

As a concept, approximating the expert's posterior belief by a non-parametric function should not differ from being willing to accept a particular choice of a prior distribution or a likelihood function. In fact, I would go as far as to say that, apart from the analytic convenience, it may be more difficult to rationalize a specific choice of prior than the expert's posterior. Prior distributions are often chosen out of convenience, but also because of their frequentist properties, or other properties such as group invariance. However, it is not clear how to translate the prior knowledge and its associated uncertainty into a probability statement when no data is at hand. In practice, we are not very concerned with this because, if the data and the model are in agreement even for a moderate sample size, the choice of a prior will have a negligible effect on the posterior distribution. Agreeably, it is impossible for an ordinary person to assign consistent probabilities across all partitions of the sample space, regardless of whether we are talking about priors, sampling density, or posteriors. So, when parametric models seem to provide a reasonable fit to the data, we can make almost uncompromising choices. A most elegant form of likelihood are those functions that can be written as log linear in the parameters, namely, the exponential family. Along with the exponential family come their useful tools such as sufficient statistics, unbi-

ased estimators, minimum variance estimators, conjugate priors, etc. This is a large enough family to cover a good portion of statistical problems. Monahan and Boos (1992) go as far as claiming that the likelihood (for Bayesian inference) must be a density of a statistic. These statements inherently assume a good degree of symmetry between the *data* and θ.

When the data and the model are not in agreement, the posterior may very well be sensitive to the choice of a prior, or some parameters may not be well identified. In that case, the posterior distribution is useless because it reflects no real uncertainty. Our personal feeling is that in modern day applications, where increasingly complex issues are modeled using statistical tools, such situations occur more often than we are willing to admit. Along the line of this thinking, examples in Dempster (1963) and the discussion in Bayarri *et al.* (1988) highlight the doubt about what, in fact, should be considered as a model for the data.

Fisher (1930) coined the term "likelihood", and he defined it to be proportional to the joint density of the observed quantities. He called it a "Mathematical Likelihood" and defined it as a measure of rational belief in the problems of parameter estimation:

> ... In the problem of estimation we start with a knowledge of the values of one or more parameters which enter into this form which values would be required for the complete specification of the population; or in other words, for the complete specification of the probabilities of the observable occurrences which constitute our data. The probability of occurrence of our entire sample is therefore expressible as a function of these unknown parameters, and the likelihood is defined merely as a function of these parameters proportional to this probability. The likelihood is thus an observable property of any hypothesis which specifies the value of the parameters of the population sampled. *Fisher(1935)*

Fisher suggests, indirectly, that to express a *rational belief* about the population distribution, we must be able to parameterize the problem in such a way that the parameter of inference, θ, completely determines the distribution of the observables. Namely, the likelihood of θ must be proportional to $P(x|\theta)$ when viewed as a function of θ. Nevertheless, Fisher provides no solution to problems where the parameters of interest do not completely specify the population distribution, which is the framework in which our example resides. The problem of estimating M is still a valid experiment, and, hence, it is a valid statistical question; we are trying to learn about an unobservable random variable M, which is best identified by the temperature profile in the opinion of the expert.

Jeffreys (1961, p.28) speaks of likelihood in more general terms: the likelihood is the probability $P(p|q_r, H)$ where p is the new information, q_r is the hypothesis under consideration, and H is the information already available. The knowledge about the current hypothesis ($M = d$) is updated via the Bayes rule by the information contained in the new data p. However, he says nothing about how to obtain $P(p|q_r, H)$ or what it should be. Of course, the posterior distribution about the hypothesis will be dependent on the knowledge of P.

On the other hand, a discussion by Bayarri *et al.* (1988) illustrates that the issue of which parameters completely identify the probability of the data is not an easy one. Conditioning on different parameters produces different likelihoods. In fact, they argue that there is no unique way for distinguishing which variables should be considered as observable and which ones as parameters in the likelihood, a point that Dempster's first example clearly brings home. They also discuss that if the full Bayesian model is specified, *i.e.* the joint distribution of all random quantities in the model, it really does not matter which parameters are part of the likelihood and which ones are part of the prior. However, Dempster's second example is the example

where the joint distribution of all random quantities is specified, but there does not exist an equivalent Bayesian model. Bayesian specification requires the knowledge of $\pi(\theta)$, and as such, it is only a single example for joint specification.

Bayarri *et al.* (1988) considered an interpretation of the likelihood provided by Berger and Wolpert (1988):

> ... θ will be understood to contain all unknown features of the probability distribution.

They point out that the statement implies that the data should be a deterministic function of θ. Therefore, the only choice for θ would be a value Y' such that $Y = Y'$ with probability 1, and 0 otherwise, and, consequently, *the ultimate likelihood* would be:

$$P(Y|\theta) \propto (y|y') = \begin{cases} 1 & \text{if } y = y' \\ 0 & \text{otherwise} \end{cases}$$

However, such a likelihood is of little satisfaction. Fisher's and Berger and Wolpert's interpretation of the likelihood both allude to the fact that θ should completely specify the sampling distribution of the data. Neither of the two definitions is general enough to incorporate the vast number of fairly complex modern day applications. I enjoyed this extreme example because it demonstrates that if you knew how to specify θ so well, you would no longer have a statistical problem.

All that being said, we may not know what the likelihood should be, but we all must agree that conditioning on the n values of a $n+1$ dimensional joint distribution clearly determines a valid conditional distribution for the $(n+1)^{th}$ variable. The only question one may have is whether $h(M)$ is a proper joint distribution, and the answer to that is yes, but it is a distribution on a space of $(\triangle_1(M), \triangle_2(M), M, \mathbf{s}(M))$, which is a map of the observed space $(x(d_1), \cdots, x(d_n), M)$.

3.6 Results

Let us look at some specific examples showing how the algorithmic likelihood performs. In the top row of figure 3.7, we plotted the observed profile and the corresponding likelihood function. As a reminder, we observe a discrete realization of $x(d)$, but the oceanographer's posteriors are elicited based on the observed intervals $(d_i, d_{i+1}], \quad i = 1, \cdots, n-1$. The interval $[d_7, d_8)$ is the most likely value for M, and it is approximately two to three times more likely than the two neighboring intervals, $[d_6, d_7)$ and $[d_8, d_9)$, approximating the oceanographer's assessment of the probability distribution. In the following two rows of figure 3.7, we alter the original profile by changing the value of temperature at depth d_8. We take two extreme cases; in the middle, we change $x(d_8)$ to be equal to that of $x(d_7)$, which changes the likelihood function to give minimal weight to the previously most likely interval $[d_7, d_8)$, and it increases the likelihood of interval $[d_8, d_9)$ to be 4.3 times more likely than $[d_6, d_7)$. In the bottom row, we decreased the temperature $x(d_8)$ to be equal to $x(d_9)$. Now the interval $[d_8, d_9)$ becomes far less likely, relative to the other intervals, than it was in the top row.

In figure 3.8, we plotted some of the more ambiguous profiles and their respective likelihood functions for M. These are examples of profiles where there is more than one likely value for M. The total probability assigned to the mode values of M in panels (a), (b), and (d) of figure 3.8 is relative to the surface temperature. Empirical studies show that waters usually mix within 0.2^oC. Therefore, if a change point appears to occur at 1.5^oC lower than the surface temperature, it becomes unlikely for that mode to be M, see figure 3.8(b). The profile in figure 3.8(c) is unusual in the sense that temperatures are recorded at uneven but very dense intervals. The likelihood function remains robust, and it could possibly be smoothed if $\triangle_2(d)$ is defined as the average drop in temperature over z meters below M. We will come

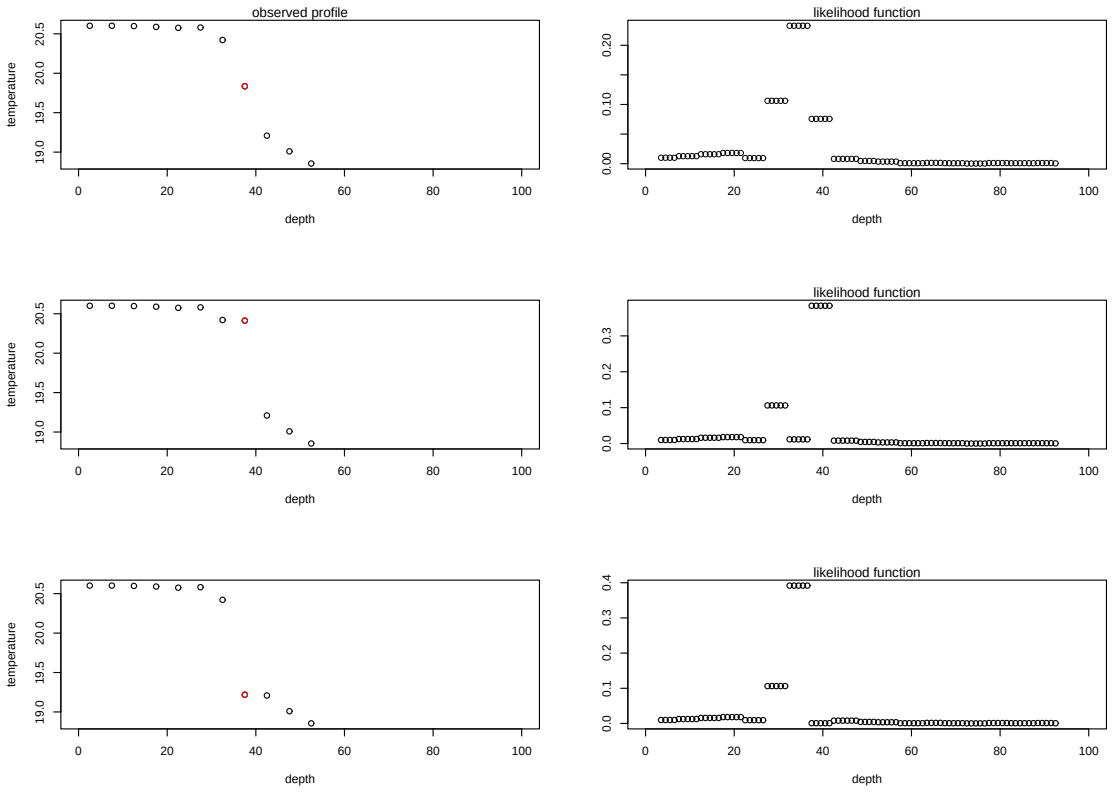


Figure 3.7: An illustration of the algorithmic likelihood

back to the same set of profiles in the next chapter. Finally, we want to give one more example that illustrates how the probability assigned to two different modal values for M is not a function of the distance between the modes, but rather of their temperatures. We took the profile from figure 3.8(c) and deleted temperatures between the depths of $90m$ and $290m$, then plotted their likelihood function in figure 3.9. Notice that the relative probability of M from approximately $30m$ to $360\,m$ stayed the same.

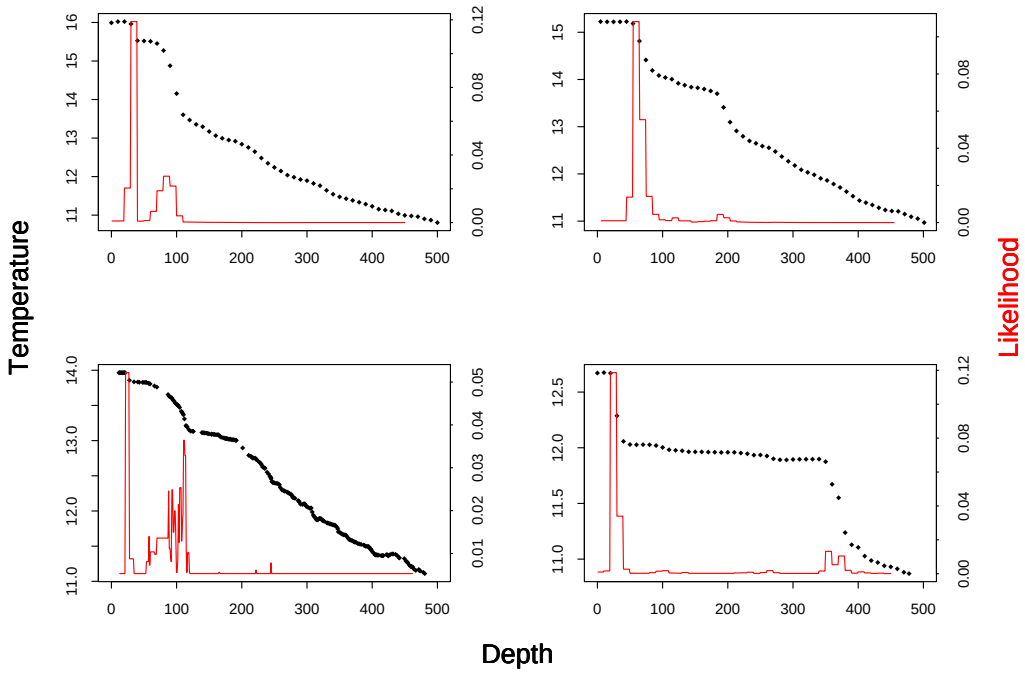


Figure 3.8: An illustration of the algorithmic likelihood

3.7 Conclusion

The goal of this section was to establish $P(data|M)$. We illustrated difficulties associated with the task. As a trade-off between simplicity and interpretability in modeling the exact (difficult and unknown) distribution, we provide an approximate likelihood elicited through the expert opinion. We invoked some uncommon ideas: equating the likelihood with the posterior distribution under the non-informative marginal density for M, and obtaining the posterior distribution from the pivotal variables. We hope that our short discussion on the literature addressing the definition of likelihood will make the reader more at ease with our novel idea.

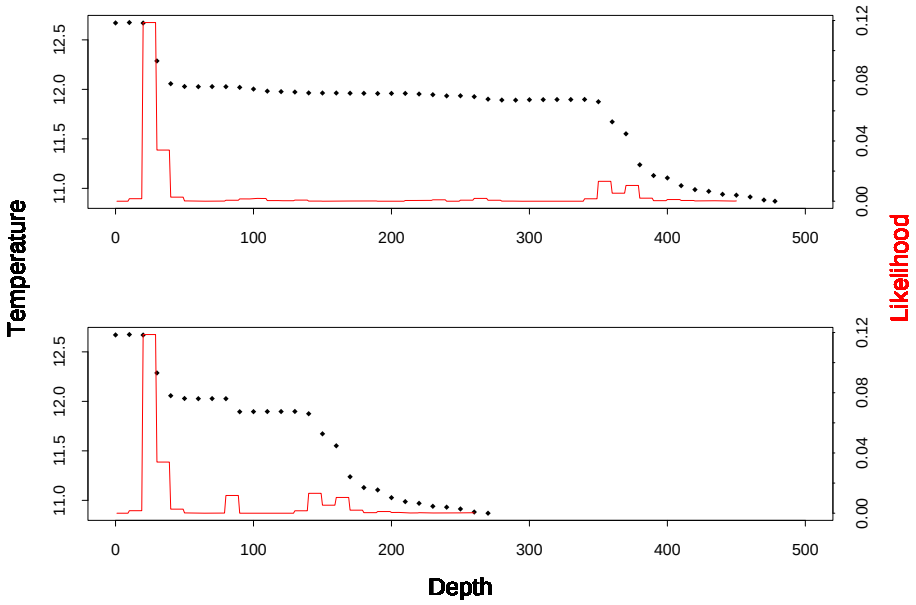


Figure 3.9: Illustration of the algorithmic likelihood on a rescaled profile

Chapter 4

Probability model for Mixed Layer Depth

4.1 Change point model for Mixed Layer depth

In this section, we will contrast the elicited likelihood with a full probability model for M at a single location in space and time. The observed thermal profile will be considered a discrete realization from a random function, $t : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, with added white noise. The shape of t is restricted to be a constant, t_s, on an interval $(0, M]$ and monotonically decreasing on (M, B), where M is a random variable and B is a previously defined depth, deeper than the Mixed layer. Function t, conditionally on (t_s, t_B, M) and rescaled by transformation, is modeled as a random distribution function via a Dirichlet Process (Lavine and Mockus, 1995). We will be concerned with one single profile at a time, and, therefore, the prior distributions for (t_s, t_B) and M are chosen as bivariate normal and uniform, respectively.

4.2 Probability model

Let $x(d_1), x(d_2), ..., x(d_n)$ be a discrete realization of temperatures at depths $d_1, d_2, ..., d_n$, and let $\epsilon(d)$ be a white noise process representing measurement error.

$$\begin{aligned} x(d) &= t_s + \epsilon(d) \text{ for } d \leq M \\ x(d) &= t(d) + \epsilon(d) \text{ for } M \leq d \leq B \end{aligned}$$

One approach in modeling the random function $t(d)$ is through a simple transformation $g : t(d) \rightarrow [0, 1]$ such that the transformation $g(t(d))$ has the properties of a cdf, and g^{-1} is readily available. One common Bayesian method for modeling a random cdf is a Dirichlet Process, DP. Conditionally on t_s and t_B, consider the transformation

$$\begin{aligned} g(t(d)) &= \frac{t_s - t(d)}{t_s - t_B} \text{ for } d \geq M \\ &= 0 \text{ otherwise,} \end{aligned}$$

and let

$$g(t(d)) \sim DP(\alpha).$$

The Dirichilet Process (DP), as characterized by Ferguson (1973), is a probability measure on an infinite dimensional space $\mathcal{P}$ of distribution functions. Consider a measurable space, $(\Theta, \mathcal{B})$, and a finite measure, α, then let a random probability measure, $\mathbf{P}$, be a stochastic process indexed by the σ-field of subsets of Θ. If, for any finite measurable partition $B_1, \cdots, B_n$ of Θ, and the distribution of the random vector $(P(B_1), \cdots, P(B_n))$ has a finite dimensional Dirichilet distribution $Dir_n(\alpha(B_1), \cdots, \alpha(B_n))$, then $\mathbf{P}$ is said to follow a Dirichlet process, $DP(\alpha)$.

The Dirichilet Process is conjugate, and some of its properties are well known. The conditional distribution of $\mathbf{P}$, given the data $X = x$, is, again, $DP(\alpha')$ where α'

is the measure defined by $\alpha'(B) = \alpha(B) + I_{\{x \in B\}}$ for each $B \in \Theta$. Additionally, the prior and posterior marginal distribution, $P(X \in B)$, are the probabilities $\alpha(B)/\alpha(\Theta)$ and $\alpha'(B)/\alpha'(\Theta)$, respectively.

The location parameter $\frac{\alpha(B_i)}{\alpha(\Theta)}$ is the prior expectation for the shape of $g(t(d))$, while $\alpha(\Theta)$ has the interpretation of a precision parameter. A large $\alpha(\Theta)$ reflects a strong prior belief in the shape of $\frac{\alpha(B_i)}{\alpha(\Theta)}$, while a small $\alpha(\Theta)$ reflects a weak prior belief. Sometimes $\alpha(\Theta)$ is thought of as a prior sample size.

Lastly, another well known result proved by Ferguson (1973), Blackwell (1973), and Sethuraman (1994)(in an alternative theorem) is that the Dirichilet Process assigns probability 1 to discrete distributions. Although, in concept, $t(d)$ is a continuous function, $x(d)$ is observed at finite depths $d_1, d_2, \cdots, d_n$. It is our feeling that modeling $t(d)$ as a DP will be adequate, in light of flexibility and ease of computation.

4.3 Distribution of the parameters

1. Distribution of errors, ϵ

   The measurement error process, $\epsilon(d)$, is considered to be Gaussian white noise with scale parameter τ_ϵ. When the roset is submerged in water, temperatures are read electronically at small intervals on a scale of just a few cm. However, because of the memory space available, only the temperatures at larger intervals, typically around 10m, are stored. Sometimes, only the average temperature of the interval is stored, which accounts for some rounding effect. Today, the instrument which measures temperature is considered precise, but measurement error occurs if the instrument is not properly maintained and calibrated. However, such errors are usually noticeable and taken care of during the data quality check at the National Oceanographic Data Center (NODC),

where data quality is performed. Data taken before the 1970s, when the electronic system was introduced, was collected by nisken bottles and temperatures were measured by hand.

With that being said, the errors are practically negligible in that their distribution is chosen for convenience, leading us to the likelihood representation:

$$\begin{aligned} P(x|t_s, t_B, \cdots) &= \prod_{i=1}^{n} \Big[P(d_i \leq M) N(t_s, 1/\tau_\epsilon) + P(d_i \geq M) N(t(d_i), 1/\tau_\epsilon) \Big] \\ &= \prod_{i\,;\,d_i \leq M} N(t_s, 1/\tau_\epsilon) \prod_{i\,;\,d_i \geq M} N(t(d_i), 1/\tau_\epsilon) \end{aligned}$$

2. Distribution of surface and bottom temperatures, t_s and t_B

   The parameters t_s and t_B are the temperatures of the Mixed Layer and depth B, respectively, and we will assume independent Gaussian distributions as their priors:

$$\begin{aligned} \pi(t_s) &\sim N(\mu_0, 1/\tau_0) \\ \pi(t_B) &\sim N(\mu_B, 1/\tau_B) \end{aligned}$$

   Parameters μ_0, μ_B, τ_0 and τ_B are elicited from the oceanographer based on the location of the profile in time and space.

3. Distribution of the Mixed Layer depth, $\mathbf{M}$

   The depth of the Mixed Layer, M, in this section will be assigned a uniform distribution:

$$\pi(M) \sim \mathcal{U}(0, B)$$

4. Distribution of the temperature profile, $\mathbf{t}(d)$

   Conditionally on t_s, t_B and M, the transformed mean temperature function, $g(t(d))$ for $d \geq M$, is a random function distributed according to a Dirichlet Process with a location- scale parameter $\alpha/||\alpha||$ or $DP(\alpha/||\alpha||)$. In this example, we express the oceanographer's prior belief and knowledge through these location and scale parameters of the Dirichlet Process. As discussed in the introductory chapter, it is widely believed that temperatures cool rapidly immediately below M and asymptote at $2°C$; therefore, we choose measure α to be:

$$\alpha((a,b)) \propto \int_a^b \lambda \, \exp(\, -\lambda \, s \,) \, ds.$$

   The hyper-parameter λ corresponds to k/ω in figure 2.3. Notice that none of the k/w values approximate the observed temperatures well, but because we want to emphasize the curvature around M, we choose λ to be 0.01.

   A prior precision parameter, $\alpha((M,B])$ or $||\alpha||$, must also be elicited. In the context of the model, a larger $||\alpha||$ will reflect a strong a priori belief in the appropriateness of the theoretical approximation, while a smaller $||\alpha||$ will reflect weaker a priori belief. Hence, a small $||\alpha||$ will be sensitive to the small observed temperature perturbations in the mixed layer. We took a training sample and asked the oceanographer to match her posterior probability with posteriors of M for various values of $||\alpha||$. Overall, $||\alpha|| = 100$ gave satisfying answers. A Dirichilet Process apriori centered around the theoretical approximation will provide us with enough flexibility to include some prior beliefs, and, at the same time, be very adaptable to the data.

5. Distribution of precision parameters $\tau_\epsilon, \gamma, \tau_0, \tau_B$

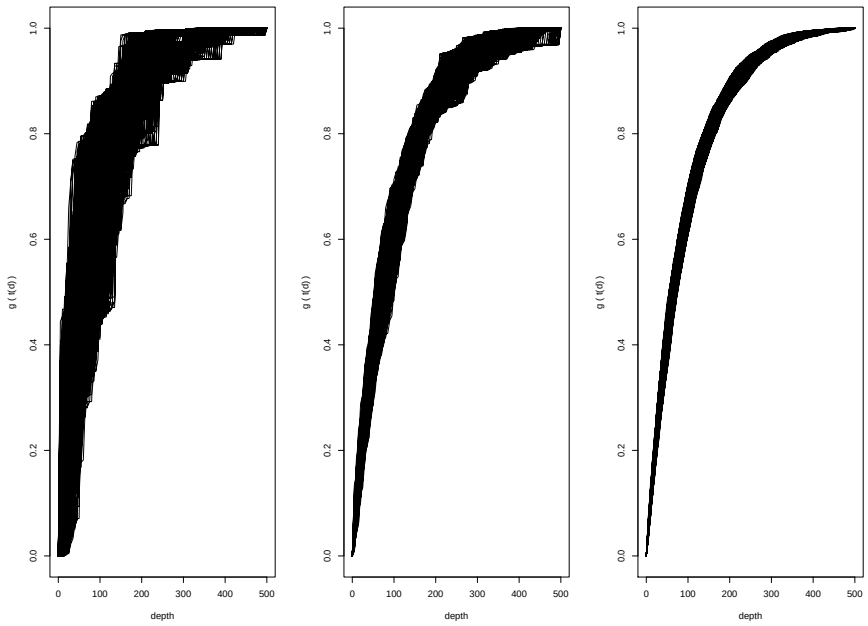


Figure 4.1: An illustration of prior precision parameter $||\alpha|| = \{10, 100, 1000\}$

The precision parameter τ_ϵ is given a conjugate $\Gamma(a_\epsilon, b_\epsilon)$ prior. The precision parameters τ_0 and τ_B are elicited from the oceanographers. It is expected that the surface temperature, t_s, at a given location in space and time, should have greater variability than those at depth B; the values are set to 1 and 10, respectively.

4.4 Sampling

Posterior distributions will be explored through a Gibbs sampler. Some conditional distributions cannot be expressed analytically. In which case, we rely on the Metropolis-Hastings step.

We will denote mean temperatures by $t_s = t(d_1) \leq, \cdots, \leq t(d_n) = t_B$. Let $i = \min\{\, j;\ d_j \geq M\}$,then $\alpha_i = \alpha((M, d_i])$ and $\alpha_j = \alpha((d_{j-1}, d_j])$, $\forall j > i$, and 0

otherwise. Let $t_{-j} = (t_1, \cdots, t_{j-1}, t_{j+1}, \cdots, t_n)$.

1. The conditional distribution $[\, t(d_2), \cdots, t(d_{n-1}),\, M | x, t_s, t_B, \tau_\epsilon, \tau_0, \tau_B \,] \propto [\, M | \cdots] \times$ $[\, t(d) | M, \cdots]$

$$[\, t(d),\, M \mid \mathbf{x}, t_s, t_B, \tau_\epsilon, \tau_0, \tau_B \,] \propto$$

$$\prod_{j=2}^{i-1} \frac{\tau_\epsilon}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}\tau_\epsilon (x_j - t_s)^2\} \prod_{j=i}^{n-1} (t_j | t_{j-1}, t_{j+1}, M, t_s, t_B) \times$$

$$\frac{\tau_\epsilon}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}\tau_\epsilon (x_j - t_n)^2\} \, \pi(M)$$

From the Dirichilet Process prior, the conditional distribution of a random vector $(g_2 - g_1,\ g_{i+1} - g_i,\ \cdots,\ g_n - g_{n-1})$ is $Dir_{n+1-i}\,(\alpha_i,\ \cdots,\ \alpha_n)$. Hence,

$$\frac{g_j - g_{j-1}}{g_{j+1} - g_{j-1}} | g_{-j} \ = \ \frac{t_{j-1} - t_j}{t_{j-1} - t_{j+1}} | t_{-j} \ \sim \ \text{Beta}(\alpha_j,\ \alpha_{j+1})$$

Therefore, the conditional distribution

$$P(\ t_j \mid M, t_{j-1}, t_{j+1}\ ) \times \frac{\tau_\epsilon}{\sqrt{(2\pi)}} \exp\{-\frac{\tau_\epsilon}{2}(x_j - t_j)^2\}$$

is a product of rescaled Beta and Gaussian densities.

For a possible, alternative error distribution, the above conditional distribution could be sampled by exact methods through successive substitutions; otherwise, a Metropolis step is required. We are interested in the distribution $t|...$ only to explore the posterior distribution of M. Since M induces support for the random function t and, consequently, defines the dimension of the parameter space, the ordinary Metropolis step does not satisfy Markovian requirements of MCMC schemes. Instead, we will consider joint sampling of M and t using a version of Reversible Jump MCMC (Green (1995), and Richardson and Green

(1997)). Suppose we have a countable collection of models $\{\mathcal{M}_k, k \in K\}$, each with its own set of parameters $\theta^{(k)} \in \mathbb{R}^{n_k}$. Let $y = (\theta^{(k)}, k)$, $\mathcal{C}_k = \{k\} \times \mathbb{R}^{n_k}$. Richardson and Green (1997) propose the construction of a Markov transition kernel, such that the probability of moving from A to B is equal to that from B to A for all A, B Borel sets in $\mathcal{C}$ by 1:1 "dimension matching" in the proposal distribution. The idea is to conceive a joint proposal distribution, $\pi(dy)q(y, dy')$, and a density with respect to some symmetric measure ξ on $\cup\mathcal{C}_k \times \cup\mathcal{C}_k$. Then the acceptance probability α is the largest when we let:

$$\alpha(y, y') = \min\left\{1, \frac{\pi(dy')\ q(y', dy)}{\pi(dy)\ q(y, dy')}\right\}$$

In particular, "dimension matching" is accomplished by proposing a value for an auxiliary variable u' which has a deterministic relationship with θ' but independent of current values of θ. Furthermore, a bijection between pairs (u, θ) and (u', θ') must exist. Then the acceptance probability will be largest when taken to be

$$\alpha(y, y') = \min\left\{1, \frac{\pi(\theta', k'|x)P(\theta'|k')q(u')}{\pi(\theta, k|x)P(\theta|k)q(u)}\left|\frac{\partial(\theta')}{\partial(\theta, u)}\right|\right\}$$

It should be noted that the Jacobian matrix and "dimension matching" in Green(1995) arises solely from introducing the auxiliary variables u and u'. Furthermore, it is easy to interpret u and u' in the context of a discrete number of change points, but it is not clear what the relationship should be between u and t, as in our example. In the discussion section of Richardson and Green(1997), Besag gives a different perspective on Reversible Jump. Suppose you have an index variable, k, along with parameter vector $\theta^{(k)}$ that defines the collection of target distributions $\pi(\theta^{(k)}, k|x)$. If the "universal target distribution" (Besag,

1987) is defined as:

$$p(k,\theta) = p(k)\pi(\theta^{(k)},k|x)\prod_{j\neq k} f_j(\pi(\theta^{(j)}))$$

where f_j is some distribution with respect to the same measure as $\pi(\theta^{(k)},k|x)$, then the pair (k,θ) lies in the space of the fixed dimension, and there is no need for introducing auxiliary variables. It is also not necessary to be able to evaluate f_j because we can construct a proposal such that the f_j terms cancel out. Similar ideas have been proposed independently by Godsill (1998) and Dellaportas *et al.* (2002).

Let the current model be $k(M)$, and define $\theta^{(k)} = t_M = t_i^M \cdots t_n^M$, $\theta = t_M \cup t_{-M}$. Then, let the proposed model be $k(M^*)$, and define $\theta^{(k^*)} = t_M^* = t_{i^*}^* \cdots t_n^*$ such that $\theta^* = t_M^* \cup t_{-M}^*$. Let t_{-M} denote all $t_{M'}(d)$ such that $M' \neq M$. Then the acceptance probability α is:

$$\begin{aligned}
\alpha &= \frac{\pi(M^*,t_M^*|y)\pi(t_{-M}^*|t_{M^*}^*,M^*)\times q(M,t|M^*,t^*)}{\pi(M,t_M|y)\,\pi(t_{-M}|t_M,M)\times q(M^*,t^*|M,t)} \\
&= \frac{\pi(M^*,t_{M^*}^*|y)\pi(t_{-M}^*|t_{M^*}^*,M^*,y)\times q_1(M)\;q_2(t_M|M,t_M^*)\;q_3(t_{-M}|t_M,M)}{\pi(M,t_M|y)\pi(t_{-M}|t_M,M,y)\times q_1(M^*)\;q_2(t_M^*|M^*,t)\;q_3(t_{-M}^*|t_{M^*}^*,M^*)} \\
&= \frac{\pi(M^*,t_M^*|y)\times\;q_1(M)\;q_2(t_M|M,t_{M^*}^*)}{\pi(M,t_M|y)\times\;q_1(M^*)\;q_2(t_M^*|M,t)},
\end{aligned}$$

with the second equality holding, if we choose $q_3(t_{-M}^*|t_{M^*}^*,M^*) = \pi(t_{M^*}^*|t_{-M^*}^*,M^*,y)$. Notice that we only need to make the choice in concept and never actually sample from this distribution.

Choose the proposal densities

$$\begin{aligned} q_1(M) &\propto c\, 1_{M \in (0,B)} \\ q_2(\, t^* \,|M, t^*, y) &= \prod_{j=n}^{i^*} q_{2_j}(t_j^* | t_{j-1}, t_{j+1}^*, M, y) \end{aligned}$$

where for each j, $q_{2_j}(t_j^*|t_{j-1}, t_{j+1}^*, M, y)$ has the density of a normal random variable $N(x_j, 1/\tau_\epsilon) 1_{t_j^* \in (t_{j-1}, t_{j+1}^*)}$.

The acceptance probability can be calculated as:

$$\begin{aligned} \alpha &= \min \Bigg\{ 1, \frac{\exp\{-\frac{1}{2}\tau_\epsilon \sum_{j=1}^{i^*-1}(x_j - t_j^*)^2 - \frac{1}{2}\tau_\epsilon \sum_{j=i^*}^{n-1}(x_j - t_j^*)^2\}}{\exp\{-\frac{1}{2}\tau_\epsilon \sum_{j=1}^{i-1}(x_j - t_j)^2 - \frac{1}{2}\tau_\epsilon \sum_{j=i}^{n-1}(x_j - t_j)^2\}} \\ &\times \frac{\prod_{j=i^*-1}^{n-1}[t_j^* - t_{j+1}^*]^{\alpha_j - 1}}{[t_s - t_B]^{\sum_{j=i^*-1}^{n-1}\alpha_j - 1}} \times \left[\frac{\prod_{j=i-1}^{n-1}[t_j - t_{j+1}]^{\alpha_j - 1}}{[t_s - t_B]^{\sum_{j=i-1}^{n-1}\alpha_j - 1}} \right]^{-1} \\ &\times \frac{\Gamma(\sum_{j=i^*-1}^{n-1}\alpha_j)}{\prod_{j=i^*-1}^{n-1}\Gamma(\alpha_j)} \times \left[\frac{\Gamma(\sum_{j=i-1}^{n-1}\alpha_j)}{\prod_{j=i-1}^{n-1}\Gamma(\alpha_j)} \right]^{-1} \\ &\times \frac{\sqrt{(}2\pi)^{-(n-i+1)}\tau_\epsilon^{n-i+1}\exp\{-\frac{1}{2}\tau_\epsilon \sum_{j=i}^{n}(t_j - x_j)^2\}}{\sqrt{(}2\pi)^{-(n-i^*+1)}\tau_\epsilon^{n-i^*+1}\exp\{-\frac{1}{2}\tau_\epsilon \sum_{j=i^*}^{n}(t_j^* - x_j)^2\}} \\ &\times \frac{\prod_{j=i-1}^{n-1}[\Phi(t_{j-1}^*) - \Phi(t_{j+1})]^{-1}}{\prod_{j=i^*-1}^{n-1}[\Phi(t_{j-1}) - \Phi(t_{j+1}^*)]^{-1}} \Bigg\} \end{aligned}$$

2. $[\, t_s, t_B | \cdots]$

The conditional distribution of $t_s, t_B|\cdots$ is also not available in the closed form:

$$
\begin{aligned}
&[t_s, t_B|\cdots] \propto \\
&\quad \exp\{-\frac{\tau_\epsilon}{2}\sum_{j=1}^{i-1}(x_j - t_j)^2\}\ \exp\{-\frac{\tau_0}{2}\ (t_s - \mu_0)^2\} \\
&\quad \exp\{-\frac{\tau_B}{2}\ (t_B - \mu_B)^2\} \exp\{-\frac{\tau_\epsilon}{2}\ (x_n - t_B)^2\} \\
&\quad \times \left[\frac{t_i - t_{n-1}}{t_s - t_B}\right]^{\sum_{j=i}^{n-1}\alpha_j} \times \left[\frac{t_s - t_i}{t_s - t_B}\right]^{\alpha_{i-1}-1} \times \left[\frac{t_{n-1} - t_B}{t_s - t_B}\right]^{\alpha_n - 1}
\end{aligned}
$$

In this case, the proposal distribution can be any bivariate distribution. However, the acceptance rate will depend on how closely the proposal distribution approximates the target distribution. Notice that the density on the first two lines of the above equation is easily calculated as a product of two independent Gaussian distributions:

$$
N\left(\frac{\tau_0\mu_0 + \tau_0\sum_{j=1}^{i-1} x_j}{\tau_0 + (i-1)\times\tau_\epsilon}, \frac{1}{\tau_0 + (i-1)\times\tau_\epsilon}\right) \times N\left(\frac{\tau_B\mu_B + \tau_\epsilon x_n}{\tau_B + \tau_\epsilon}, \frac{1}{\tau_B + \tau_\epsilon}\right)
$$

As a proposal distribution we use truncated normals:

$$
\begin{aligned}
t_s^* &\sim N(\ \mu_1, \sigma_1^2;\ t_s^* > t_i) \\
t_b^* &\sim N(\ \mu_2, \sigma_2^2;\ t_b^* < t_{n-1})
\end{aligned}
$$

which leads to the acceptance probability:

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\exp\{-\frac{\tau_\epsilon}{2}\sum_{j=1}^{i-1}(x_j - t_s^*)^2 - \frac{\tau_\epsilon}{2}(x_n - t_B^*)^2\}}{\exp\{-\frac{\tau_\epsilon}{2}\sum_{j=1}^{i-1}(x_j - t_s)^2 - \frac{\tau_\epsilon}{2}(x_n - t_B)^2\}}\right. \\
&\times \frac{\exp\{-\frac{\tau_0}{2}(t_s^* - \mu_1)^2 - \frac{\tau_B}{2}(t_B^* - \mu_2)^2\}}{\exp\{-\frac{\tau_0}{2}(t_s - \mu_1)^2 - \frac{\tau_B}{2}(t_B - \mu_2)^2\}} \\
&\times \left[\frac{t_i - t_{n-1}}{t_s^* - t_B^*}\right]^{\sum_{j=i}^{n-1}\alpha_j} \times \left[\frac{t_s^* - t_i}{t_s^* - t_B^*}\right]^{\alpha_{i-1}-1} \times \left[\frac{t_{n-1} - t_B^*}{t_s^* - t_B^*}\right]^{\alpha_n - 1} \\
&\times \{\left[\frac{t_i - t_{n-1}}{t_s - t_B}\right]^{\sum_{j=i}^{n-1}\alpha_j} \times \left[\frac{t_s - t_i}{t_s - t_B}\right]^{\alpha_{i-1}-1} \times \left[\frac{t_{n-1} - t_B}{t_s - t_B}\right]^{\alpha_n - 1}\}^{-1} \\
&\times \left.\frac{\exp\{-\frac{1}{2\sigma_1^2}(t_s^* - \mu_1)^2 - \frac{1}{2\sigma_2^2}(t_B^* - \mu_2)^2\}}{\exp\{-\frac{1}{2\sigma_1^2}(t_s - \mu_1)^2 - \frac{1}{2\sigma_2^2}(t_B - \mu_2)^2\}}\right\}
\end{aligned}$$

3. The conditional distribution for the precision parameter $[\,\tau_\epsilon|\cdots\,]$

   Since the precision parameter has been assigned conjugate priors, its conditional distribution is analytical and can be sampled within the Gibbs algorithm. The full conditional distribution for the precision parameter is:

$$\tau_\epsilon|\cdots \;\sim\; InvGa\left(a + \frac{n}{2}, b + \frac{1}{2}[\,\sum_{j=1}^{i-1}(x_j - t_s)^2 + \sum_{j=i}^{n}(x_j - t_j)^2\,]\right).$$

4.5 Results and Discussion

In this section, we will show results for the posterior distribution of M and compare it to the performance of the algorithmic likelihood introduced in the previous chapter.

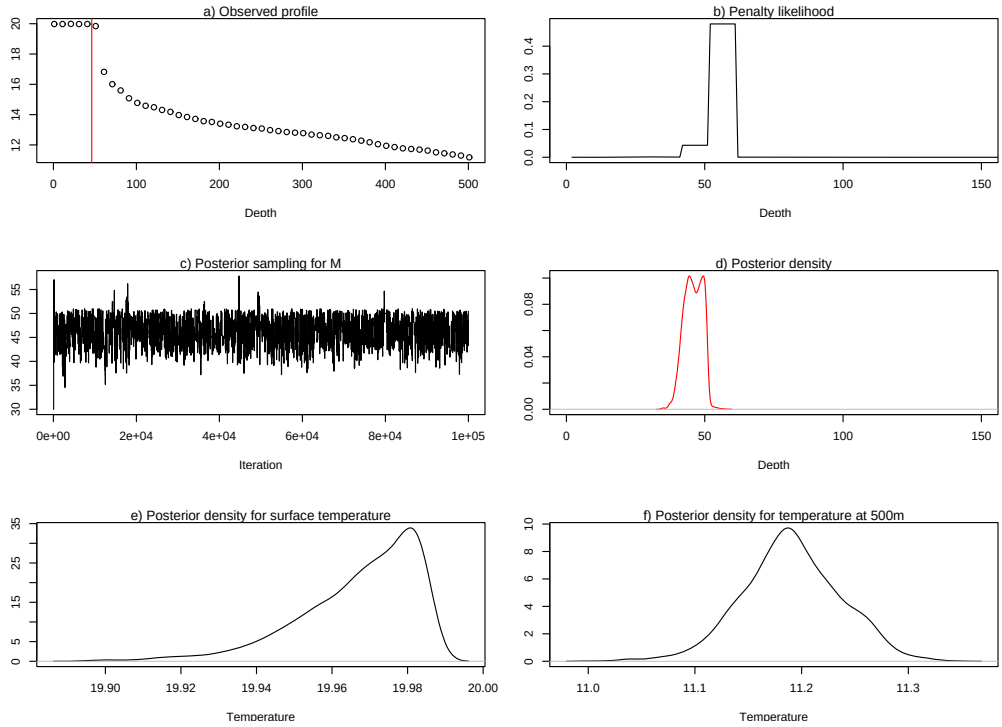


Figure 4.2: a) observed profile, b) algorithmic likelihood, c) MCMC trace-plot for the probability model d) posterior distribution for M e) posterior density estimate for the surface temperature f) posterior density estimate for the temperature at $500m$

Consider the profile in figure 4.2. Because of the large difference in temperatures between d_6 and d_7, the oceanographer concludes, with high certainty, that the waters are mixing to the depth of d_7. She also assigns a 1 in 10 chance that the depth of the Mixed Layer is in the interval $[d_5, d_6)$ since $x(d_6)$ is cooler than the surface temperature. This profile identifies M clearly, and the posterior probability is de-facto split between these two intervals in the ratio of $1 : 10$. The posterior probabilities, according to the algorithmic likelihood, and the change point model are found in panels (b) and (d) of 4.2. Panel (c) of 4.2 shows the MCMC trace plot of the samples from the posterior. The chain mixes well, but a great deal of posterior mass is on

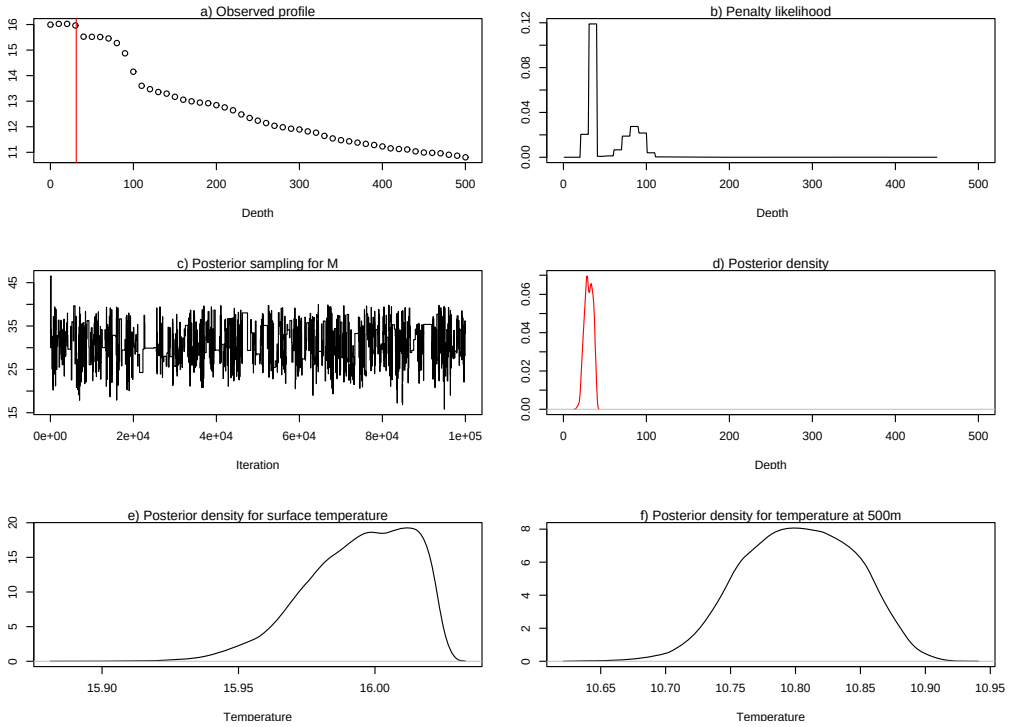


Figure 4.3: a) observed profile, b) algorithmic likelihood, c) MCMC trace-plot for the probability model d) posterior distribution for M e) posterior density estimate for the surface temperature f) posterior density estimate for the temperature at $500m$

the interval $[d_5, d_6)$ rather than $[d_6, d_7)$. The explanation resides in the choice of prior precision parameter $||\alpha||$ which favors the change point at even small drops of temperatures below M. A different choice of $||\alpha||$ may favor different intervals. In particular, a larger $||\alpha||$ would give more probability to interval $[d_6, d_7)$, but it also may have an effect on the mixing of the chain. Previously, we commented that the choice of α had a relationship to a thoretical approximation. However, there is no particular justification for the value of $||\alpha||$.

Next, let us consider the other three profiles previously discussed in Chapter 3, now plotted in figures 4.3, 4.4 and 4.5. In all three of these profiles, the oceanographer

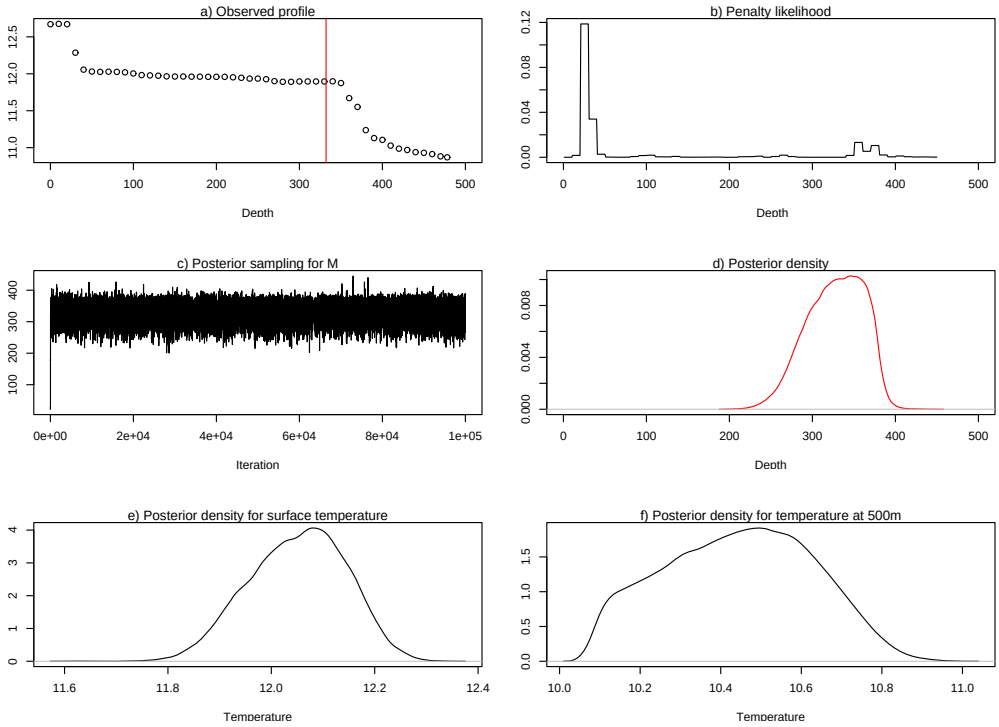


Figure 4.4: a) observed profile, b) algorithmic likelihood, c) MCMC trace-plot for the probability model d) posterior distribution for M e) posterior density estimate for the surface temperature f) posterior density estimate for the temperature at $500m$

makes an assessment that there are three distinct modes of posterior distribution. She divides her belief according to the difference in temperatures from the surface and the rate of cooling immediately below the depth M. In all three examples, the change point model produces a unimodal posterior, which is due to the choice of the likelihood and L_2 loss function in the estimation of M. We will use a likelihood ratio statistic in two extreme cases to illustrate the point. Consider the profile in figure 4.4:

$$\begin{aligned} \text{Let current } M &\in (d_3, d_4] \text{ and let } t = (x_1, x_1, x_1, x_{36}, \cdots, x_{49}) \\ \text{Propose } M^* &\in (d_{36}, d_{37}] \text{ and let } t^* = (x_5, \cdots, x_5, x_{37}, \cdots, x_{49}) \end{aligned}$$

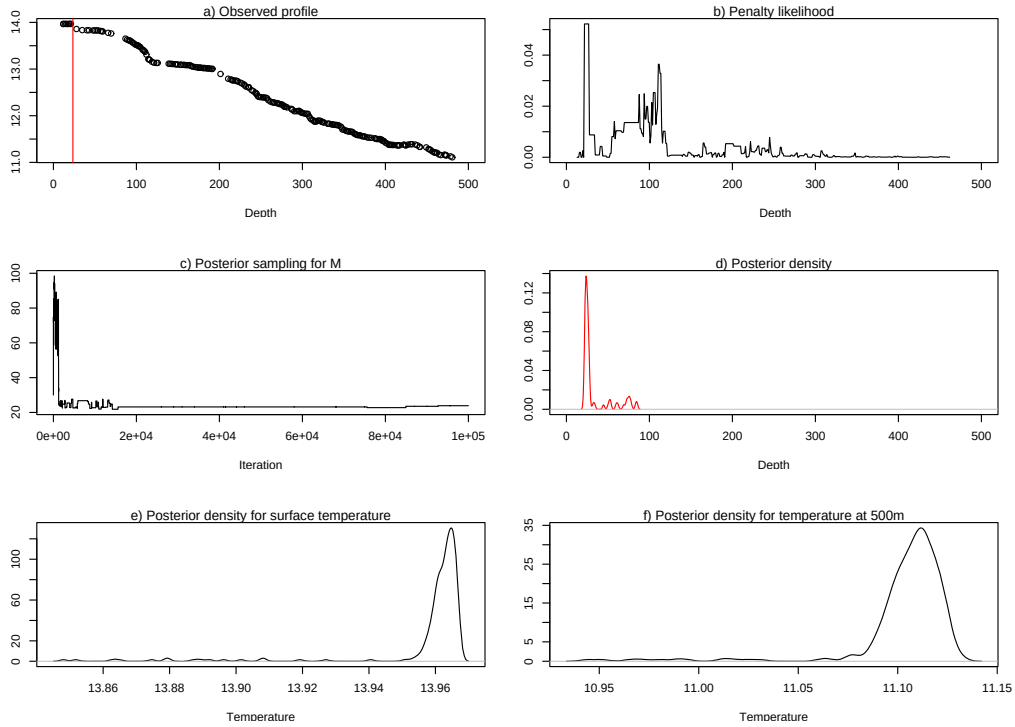


Figure 4.5: a) observed profile, b) algorithmic likelihood, c) MCMC trace-plot for the probability model d) posterior distribution for M e) posterior density estimate for the surface temperature f) posterior density estimate for the temperature at $500m$

Then the likelihood ratio is

$$= \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t(d_i)^{*}-x(d_i))^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t(d_i)-x(d_i))^2}},$$

which is approximately equal to

$$c\; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t(d_i)^{*}-t(d_i))^2},$$

and as $n \to \infty$, the likelihood ratio is roughly

$$c\; e^{-\frac{1}{2\sigma^2}\times(d_4-d_1)*(x(d_1)^{-}x(d_5))^2} < \infty$$

Therefore, the likelihood ratio is directly dependent on the distance from the surface to the first modal value for M, *i.e.* the distance $(d_4 - d_1)$.

Now, consider the same current model and propose

$$M^* \in (d_{36}, d_{37}] \text{ and } t^* = (x_1, \cdots, x_1, x_{37}, \cdots, x_{49})$$

Then the likelihood ratio is

$$= \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t(d_i)^* - x(d_i))^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t(d_i) - x(d_i))^2}},$$

which is approximately equal to

$$c\, e^{-\frac{1}{2\sigma^2}\sum_{i=4}^{36}(x(d_1) - x(d_i))^2}.$$

Once again, as $n \to \infty$, the likelihood ratio is roughly

$$c\, e^{-\frac{1}{2\sigma^2} \times (d_{36} - d_1) * (x(d_1)^- x(d_{36}))^2}.$$

Again, this is directly dependent on the distance between the modes. As two modes move further apart, it becomes more difficult to capture both of them using this model. As an illustration, we deleted observations between $100m$ and $300m$ (as in chapter 3) from the profile in figure 4.4. When the oceanographer estimates her posterior belief, she distributes the probability to the modes according to the change in temperature from the surface and the rate of decay immediately below M. Consequently, her posterior does not change, it only shifts the second mode by 200 deleted meters (see figure 4.6). The resulting posterior is now bimodal with mass on the two modes: one narrow in the range of $10 - 25$ meters and the other in the range of $90 - 200$ meters.

A more practical issue of computational efficiency must also be considered. The overall acceptance rate in these computations varied between .1% and 10%, so it

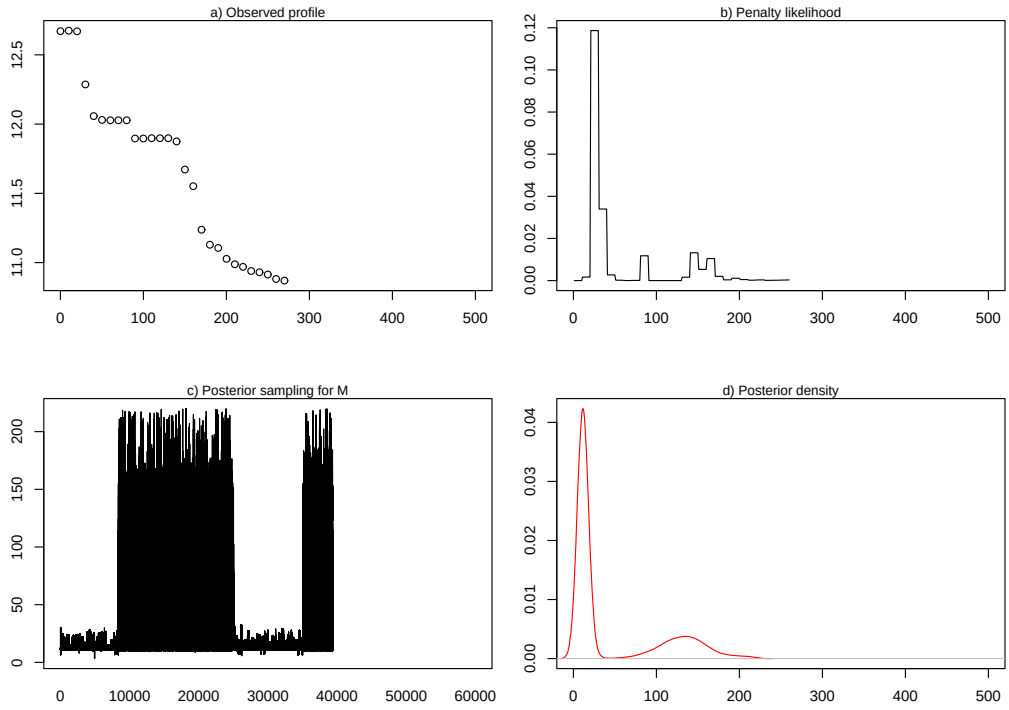


Figure 4.6: a) observed profile, b) algorithmic likelihood, c) MCMC trace-plot for the probability model d) posterior distribution for M

is necessary to run the chain for many iterations. We see that the four chains in figures 4.2(c), 4.3(c), and 4.4(c) all appear to mix reasonably well. Because it is very difficult to move the chain out of a saddle point, we may see behavior similar to that in figure 4.5(c). It may be acceptable if it happens every once in a while, but it would become infeasible to monitor the chains for each profile separately when the model is extended to the spatio-temporal domain.

Chapter 5

Space time model

5.1 Spatio-temporal model

As discussed in the introduction, the season's annual cycle forces a change in the temperature profile for each specific location in the ocean in addition to changing the depth of the Mixed Layer. Temperature profiles are also interdependent due to their proximity in space. In this section, we will introduce a spatio-temporal model for M.

We will denote the observed temperatures, $X(s,t,d) = x(s,t,d)$, as a function of yearday, space, and depth. Let $s = (s_1, s_2)$ denote latitude and longitude, respectively. It is reasonable to assume that the unobserved variable $M(s,t)$, defined by $x(s,t,\cdot)$, is a smooth function of space and time. Therefore, apriori, we say $M(s,t)$ is a noisy realization from a zero mean Gaussian Process and a long term trend whose posterior distribution will be estimated by the data. The hierarchical specification of the model is:

$$\begin{aligned} lik(\ M(s,t)\ ;\ X(s,t,\cdot)) &\propto h(\ \triangle_1(\ M(s,t)\ ),\ \triangle_2(\ M(s,t)\ )\ ) \\ M(s,t) \mid \gamma(s,t) &\sim N(\gamma(s,t),\ \sigma_\epsilon^2\, I\ ) \end{aligned}$$

where

$$\gamma(s,t) = \boldsymbol{Y}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\beta} + \mu(s_1,t)$$

and

$$\mu(s_1, t) \equiv \int k(s_1 - \omega, |t - \upsilon| \bmod C)\, \xi(\omega, \upsilon)\, d(\omega, \upsilon)$$

In the specification above, $h(M)$ is as before, $\boldsymbol{Y}$ is a matrix of population regressors, and $\boldsymbol{\alpha}$ is the vector of corresponding population parameters. Matrix $\boldsymbol{Z}$ is a design matrix, linking factors $j = 1, \cdots J$ to the response variables M, and $\boldsymbol{\beta}$ is a vector of corresponding factor-specific parameters. The process $\mu(s_1, t)$ accounts for a latitude dependent, inter-annual seasonal cycle and is modeled as a Gaussian process by convolution of a white noise latent process, ξ, and a Gaussian kernel, k. The latent process ξ has latitude and yearday coordinates ω and υ, respectively. The kernel function k is a separable, two dimensional Gaussian kernel,

$$k(s_1 - \omega, |t - \tau| \bmod C) = k^{s_1}(s_1 - \omega) \times k^t(|t - \upsilon| \bmod C),$$

with C being the circumference of a circle representing the length of a seasonal cycle (365 days). Such construction yields a periodic, inter-annual seasonal cycle $\mu(s_1, \cdot)$ smoothly varying with the latitude coordinate s_1.

On a large scale, at any given time in the ocean, temperatures, as well as M, may be regarded as realizations of random processes. As long as we stay away from the large currents and coastal regions, these quantities are envisioned to have a similar correlation structure at all spatial locations. On a smaller scale, there are other sources of variation due to external forcing such as hurricanes, trade winds, precipitation, etc. We don't have the data on such activities, but since research cruises typically last no longer than a few weeks and cover relatively small areas, we believe short term deviations may be accounted for by an extra variation parameter, β_{cruise}, one for each cruise.

Thousands of data points have been made available over the years, but still inherent in the sampling design is the difficulty distinguishing between the trend and the

noise. Expeditions typically take measurements along a direct linear course within small temporal windows, always making the information available in only one out of three dimensions. Therefore, in this study, we are concerned with exploring the relationship between the covariates and response variables rather than estimating properties such as the smoothness of the random process, variogram parameters, etc., which are difficult to estimate given the restrictions of the sampling design. By regarding the design matrix $\boldsymbol{Z}$ as a general design matrix, we can explore decadal trends, trends by different latitude and longitude bands, nonlinear trends, etc.

5.2 Convolution approach

In general terms, we say that the random process $Z(\mathbf{s})$ is a Gaussian process if, for $n \geq 1$, any finite dimensional vector, $Z(s_1, s_2, \cdots, s_n)$, is distributed as $MVN_n(\mu(\mathbf{s}), \Sigma)$. In practice, we have a single realization of a random process, and Σ is not known. Therefore, in order to estimate parameters of the mean process $\mu(s)$, further restrictions are needed. The most common restriction considered is the stationarity of the stochastic process. We say that a process is *strictly stationary* if $Z(\mathbf{s})$ satisfies: for $\{s_1, \cdots, s_n\} \in \mathbb{R}^d$ and any $\mathbf{h} \in \mathbb{R}^d$ for $d \geq 1$,

$$P(Z(s_1 + h), \cdots, Z(s_n + h)) = P(Z(s_1), \cdots, Z(s_n))$$

A less restrictive type of stationarity, referred to as *weak stationarity*, is defined by the first two moments of $Z(s)$. The process is said to be weakly stationary if it has a constant mean, $\mu(s) = \mu$, and the covariance function, $Cov(Z(s), Z(s')) = C(s - s')$, depends only on the separation vector and not the direction. Strong stationarity, by definition, implies weak stationarity. The converse is also true for Gaussian Processes. For C to be a proper covariance function, it must be a positive definite, *i.e.* it must

have the following representation: for any complex or real valued $c_1, \cdots, c_n$

$$\sum_{i,j} c_i \, c_j Cov(Z(s_i), Z(s_j)) = \sum_{i,j} c_i \, c_j C(s_i, s_j) \; \geq \; 0.$$

The most common approach to obtain a positive definite function, $C(h)$, is to take a Fourier transform:

$$C(\mathbf{h}) \; = \; \int_{\mathbb{R}^d} \exp i\omega^T \mathbf{h} F(d\boldsymbol{\omega}) \tag{5.1}$$

of a positive finite measure, F. If $F(\boldsymbol{\omega})$ is absolutely continuous (i.e. has density $f(\boldsymbol{\omega})$ with respect to the Lebesgue measure), then $f(\boldsymbol{\omega}) = F(d\boldsymbol{\omega})/d\boldsymbol{\omega}$ is called spectral density. *Bochner's Theorem* (Gihman and Skorohod (1974) pg. 208) states that, in fact, a spectral representation of a covariance function is a necessary and sufficient condition for $C(s_i - s_j)$ to be a positive definite. A convolution of any two valid covariance functions leads to yet another valid covariance function. Let C_1 and C_2 be two covariance functions with a spectral representation as in 5.1 and with spectral densities $f_1(\omega)$ and $f_2(\omega)$, respectively. Then,

$$C(\mathbf{h}) = \int C_1(\mathbf{h} - \mathbf{t}) \, C_2(\mathbf{t}) d\mathbf{t}$$

is also a covariance function since it has the following spectral representation:

$$\begin{aligned} f(\boldsymbol{\omega}) \; &= \; \int \exp\{-i\boldsymbol{\omega}^T \mathbf{h}\} \, C(\mathbf{h}) \, d\mathbf{h} && (5.2) \\ &= \; \int \exp\{-i\boldsymbol{\omega}^T \mathbf{h}\} \int C_1(\mathbf{h} - \mathbf{t}) \, C_2(\mathbf{t}) \, d\mathbf{t} \, d\mathbf{h} && (5.3) \\ &= \; f_1(\boldsymbol{\omega}) \times f_2(\boldsymbol{\omega}) \end{aligned}$$

The convolution approach for constructing a Gaussian process is based on noticing that the kernel convolution of a Gaussian white noise process

$$Z(\mathbf{h}) = \int_{\mathbb{R}^d} C_1(\mathbf{h} - \mathbf{t}) \xi(\mathbf{t}) d\mathbf{t}$$

yields exactly the covariance function

$$C(\mathbf{h}) = \int C_1(\mathbf{h}-\mathbf{t})C_1(\mathbf{t})dt$$

for the Gaussian process $Z(\mathbf{h})$. Moreover, since the inverse Fourier transform, *i.e.* the spectral density of process $Z(h)$, is a square root of known spectral density $f_1 = f_2$, it is possible that for some (the most commonly used) covariance function, $C(h)$, we compute the kernel function $C_1 = C_2$ by:

$$C_1(h) \propto \int \exp i\boldsymbol{\omega} h \sqrt{f(\boldsymbol{\omega})} d\boldsymbol{\omega}$$

For many of the important members of the Matern class of covariance functions, there is a one to one correspondence with the kernel function. The analytic form is available for exponential and Gaussian correlogram.

5.3 Model specification

Let $\mu(s_1, t)$, in latitude $\times$ yearday domain, be defined as before:

$$\mu(s_1, t) \equiv \int k(s_1 - \omega, |t - \upsilon| \bmod C)\, \xi(\omega, \upsilon)\, d(\omega, \upsilon) \tag{5.4}$$

with

$$\begin{aligned} E(\mu(s_1, t)) &= 0 \\ Corr(\mu(s_1, t), \mu(s_1^{'}, t^{'})) &= \int k^{s_1}(s_1 - \omega) \times \\ & \qquad k^t(|t^{'} - \upsilon| \bmod C)\, d(\omega, \upsilon) \end{aligned}$$

After observing the data, parameters of the process $\xi(\omega, \upsilon)$ must be estimated so that the integral in 5.4 is replaced by summation. Let h index a discretized process, ξ, in

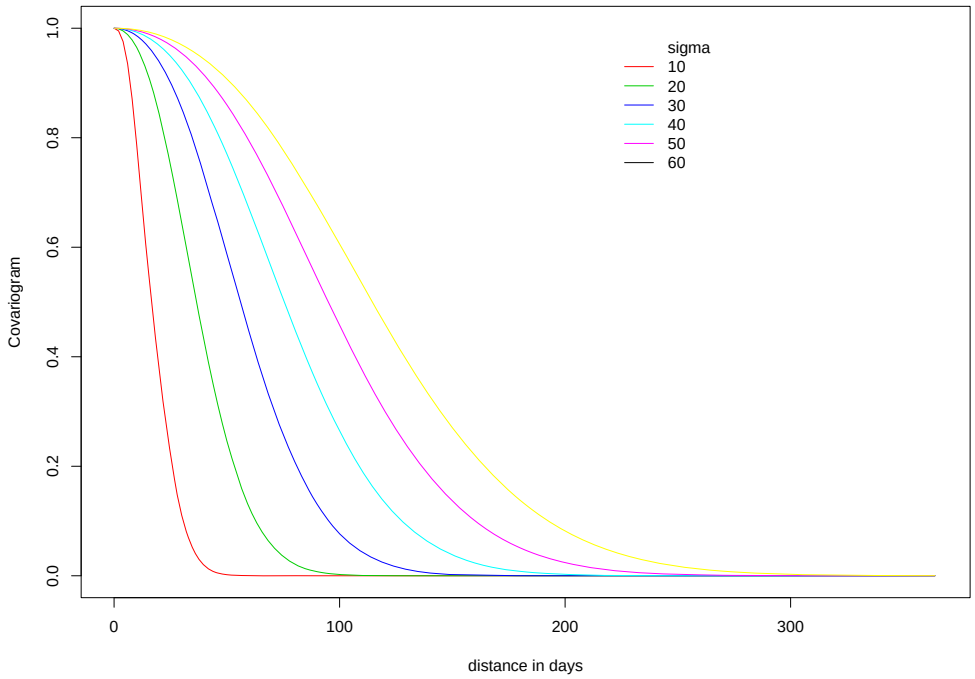


Figure 5.1: An illustration of the covariogram induced by the choice of a precision parameter in a Gaussian kernel.

latitude-yearday dimension, then

$$\mu(s_1, t) \approx \sum_{h=1}^{H} k^{s_1}(s_1 - \omega_h) k^t(|t - \upsilon_h| \bmod C)\, \xi_h \tag{5.5}$$

$$= K\, \xi \tag{5.6}$$

The entries of a kernel weight matrix K must be normalized such that the diagonal elements of the covariance $Corr(\mu(s_1, t), \mu(s_1^{'}, t^{'}))$ are unity. Now we can see how the discretization of ξ leads to computational efficiency because the number of parameters which need to be estimated depends on H rather than sample size. Similar to any smoothing procedure, care needs to be taken as to how the process is discretized. Poor choices may lead to over-smoothing or under-smoothing.

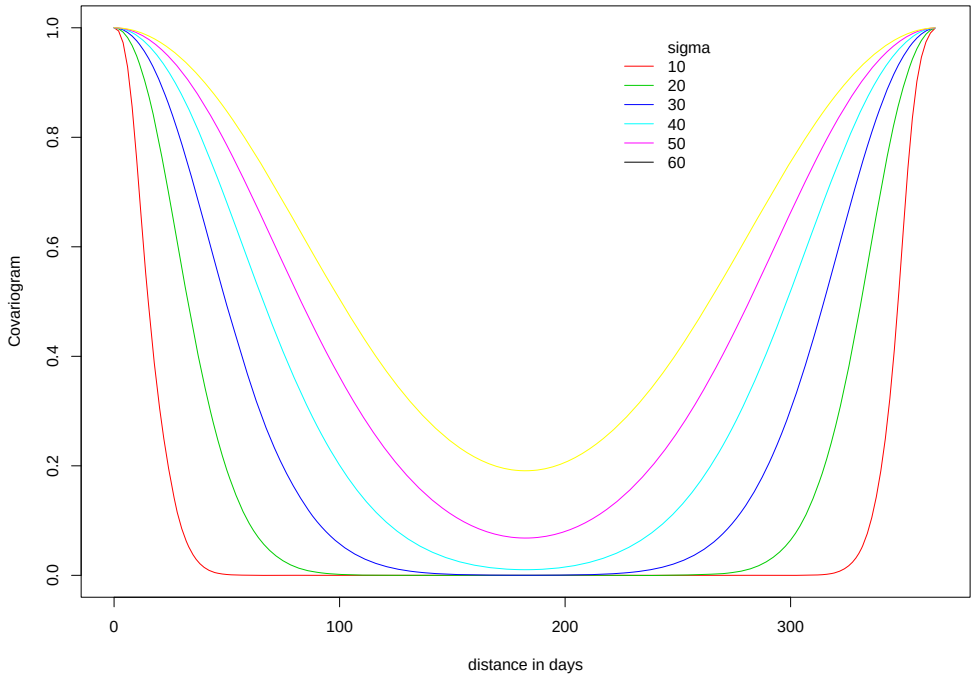


Figure 5.2: An illustration of the covariogram induced by the choice of a precision parameter in a Gaussian kernel in the temporal domain.

To complete the model specification, we choose prior distributions:

$$\begin{aligned}
\boldsymbol{\alpha} &\sim N(\mathbf{0}, \mathbf{B}_0) \\
\boldsymbol{\beta} &\sim N(\mathbf{0}, \mathbf{B}_1) \\
\xi_h &\sim N(0, \sigma_\xi^2) \qquad h = 1, \cdots, H \\
1/\sigma_\epsilon^2 &\sim \Gamma(a_\epsilon, b_\epsilon)
\end{aligned}$$

The full conditional distributions are available for all parameters except for M. Therefore, the efficient method to sample from the posterior distribution is via the Gibbs sampler with a Metropolis-Hasting step for M Gelfand *et al.* (1990),Gilks *et al.* (1993).

5.3.1 Sampling

I. Metropolis-Hastings step: Conditionally on current values of the mean process $\gamma(s,t)$, we propose a new move

$$M_i^* \sim \mathcal{U}[min(d_i), max(d_i))$$

and accept the move $M_i \rightarrow M_i^*$, with probability α, where

$$\alpha = \min\left\{1 \ , \ \frac{h(\triangle_1(M_i^*), \triangle_2(M_i^*))\exp\{-\frac{1}{2\sigma_\epsilon^2}(M_i^* - \gamma_i)^2}{h(\triangle_1(M_i), \triangle_2(M_i))\exp\{-\frac{1}{2\sigma_\epsilon^2}(M_i - \gamma_i)^2}\right\}$$

for all $i = 1, \cdots, n$.

II. The overall mean parameter $\boldsymbol{\alpha}$:

$$[\boldsymbol{\alpha}|\cdots] = N(\boldsymbol{S}^{-1}\boldsymbol{Y}^T\boldsymbol{\epsilon}_0, \boldsymbol{S}^{-1})$$

where

$$\boldsymbol{S} = \mathbf{B}_0{}^{-1} + \boldsymbol{Y}^T\boldsymbol{Y}/\sigma_\epsilon^2$$

and

$$\boldsymbol{\epsilon}_0 = \mathbf{M} - \sum_j \mathbf{Z}_j\boldsymbol{\beta}_j$$

III. Parameters $\boldsymbol{\beta}$, for each factor $j = 1, \cdots, J$, taking level $p = 1, \cdots, P$:

$$[\boldsymbol{\beta}_{pj}|\cdots] = N(\boldsymbol{S}_{pj}^{-1}\sum_{i:P(i,j)=p}\boldsymbol{Z}_{pj}\epsilon_{ij}/\sigma_\epsilon^2, \boldsymbol{S}_{pj}^{-1})$$

with the sum being over all i with factor j at level p, and

$$\boldsymbol{S}_{pj} = \mathbf{B}_1^{-1} + \sum_{i:P(i,j)=p}\boldsymbol{Z}_{ij}^T\boldsymbol{Z}_{ij}/\sigma_\epsilon^2$$

and

$$\epsilon_{ij} = M_i - \mathbf{y}_i^T\boldsymbol{\alpha} - \sum_{g:g\neq j}\boldsymbol{Z}_{jg}\boldsymbol{\beta}_{P(i,g)}$$

where the sum is over all factors g except j.

IV. The latent process variables ξ_h for $h = 1, \cdots, H$:

$$[\xi_h | \cdots] \sim N(S^{-1} K_h^T \epsilon_h^T / \sigma_\epsilon^2, S^{-1})$$

where

$$S = \frac{1}{\sigma_\epsilon^2} K_h^T K_h + \frac{1}{\sigma_\xi^2}$$

and

$$\epsilon_h = \mathbf{M} - \boldsymbol{Y}\boldsymbol{\alpha} - \boldsymbol{Z}\boldsymbol{\beta} - \sum K_{-h}^T \xi_{-h}$$

with the sum being over all $g = 1, \cdots, H$ except for h.

V. Variance parameter σ_ϵ^2:

$$[1/\sigma_\epsilon^2 | \cdots] \sim \Gamma(a_\epsilon + \frac{n}{2}, b_\epsilon + \frac{1}{2} \sum_{i=1}^{n} (M_i - \gamma_i)^2)$$

VI. Additionally, for identifiability between $\mu(s_1, t)$ and α, at each iteration the process μ is normalized such that for all $K^h, h > 0$:

$$\sum_{i=1}^{n} {K_i}^h \xi_h = 0$$

.

5.4 Analysis and results

We created three scenarios in the analysis of the ocean's Mixed Layer.

I. A common intercept and slope for the annual effect among all observations, and the design matrix $\boldsymbol{Z}$ linking each cruise to the response variable M.

II. A common intercept and $\boldsymbol{Z}$ linking the year to the response variable for four different latitude bands.

III. Latitude band or decade specific intercept term.

We applied all three scenarios to the two specified regions in the ocean: a sample from the subtropical gyre and a sample from the subpolar gyre. Additionally, to make a comprehensive statement on the influence of increased heat flux at the surface of the ocean, we applied a similar analysis to the observed temperatures on specific cross-sections of the ocean. Specifically, we conducted an analysis on temperatures at $0, 50, 100, 250$ and $500\,m$ for the subtropical gyre and on the temperatures at $0, 100$, and $500\,m$ for the subpolar gyre.

5.4.1 Subtropical gyre

There were 5,820 thermal profiles used in the analysis of the subtropical gyre. We used a Gaussian kernel with a standard deviation of $175\,km$ in space, on a grid discretized every $100\,km$ and a standard deviation of 25 days in time, with the latent grid discretized by 14 days. Refining the discretization of the latent process is limited by the available computer power.

Let us begin the analysis with what was started in scenario I: modeling the long term trend as linear in year and additive to the space-yearday process, which accounts for latitudinally dependent - seasonal variability. Posterior density for the slope and intercept are plotted with black lines in figure 5.3, indicating that the depth of the Mixed Layer, on average, is shoaling $.15m$ per year or $7.5m$ over the past five decades. We introduced an additional variable *cruise effect* which is constrained to a zero mean within each year to account for the unbalanced sampling design and short temporal scale variability associated with each cruise. The resulting posterior distribution for each parameter is found in red in figure 5.3. The mean of the intercept parameter has increased by $6m$ to $53.03m$, the amount equal to the previously estimated change over four decades, while the slope parameter has mostly changed in the variance.

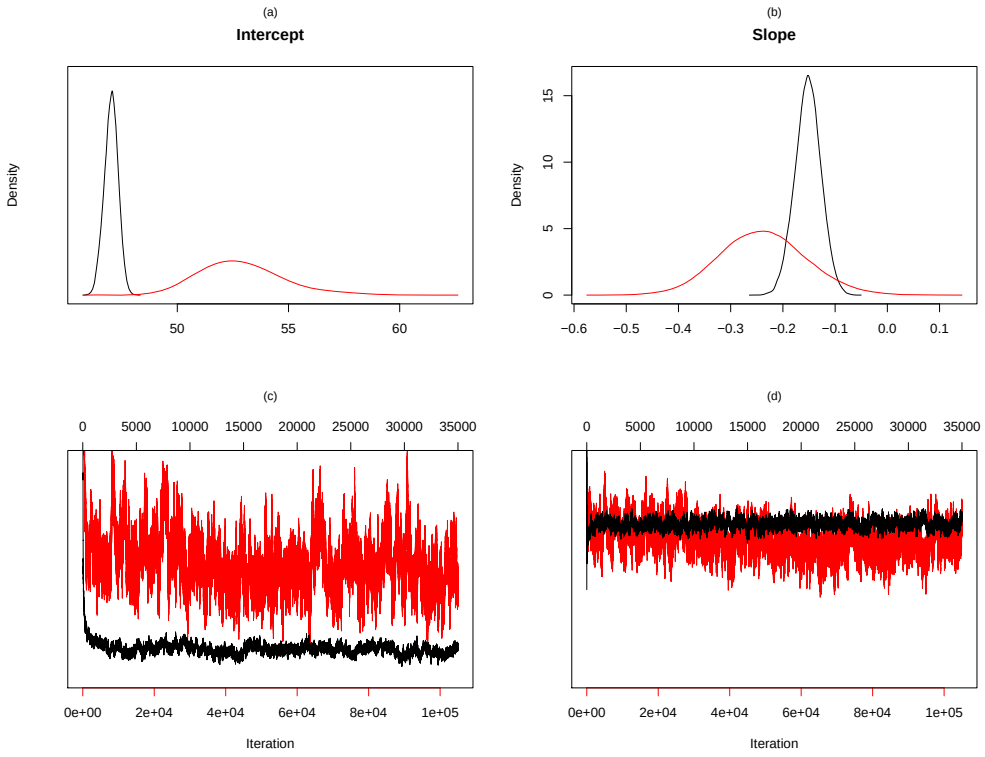


Figure 5.3: Posterior densities for the common mean and annual effect in the subtropical gyre.

The correlation between the Markov chains for slope and intercept is negligible, but the model with the *cruise effect* exhibits much more autocorrelation in the MCMC samples. In figure 5.4, we plotted the posterior means of the cruise effect vs year by latitude bands, starting with $34-38N$ in the top left panel, $29-34N$ in the top right panel, $24-29N$ in the lower left panel, and $20-24N$ in the lower right panel. Notice that at lower latitudes, more cruises have a negative contribution than positive, which explains the change in the intercept parameter. Additionally, large values of *cruise effect* are found in the months of March and April, when rapid transitions between cold and warm months take place. However, it is more important to note that by accounting for *cruise effect* the slope parameter changes to $-.25\,m/year$, and hence, M is estimated to have shallowed by $12.5\,m$ over the last five decades.

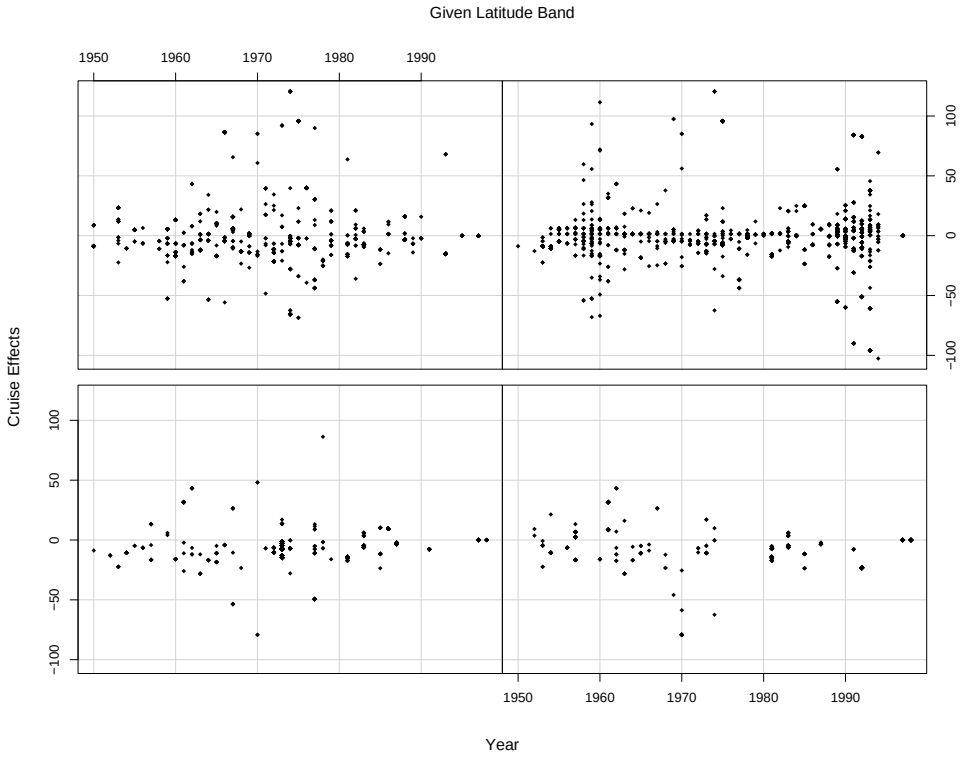


Figure 5.4: Posterior means for the *cruise effects* in the subtropical gyre given the latitude band. Top down by row: 34-38N, 29-34N, 24-29N, 20-24N.

As mentioned earlier, we performed a similar analysis for the temperatures at different depths. The model is as follows:

$$
\begin{aligned}
T(s,t,d) \mid \gamma(s,t) &\sim N(\gamma(s,t),\ \sigma_\epsilon^2 I\ ) \\
&\text{where} \\
\gamma(s,t) &= \boldsymbol{Y\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mu(s_1,t) \\
&\text{and} \\
\mu(s_1,t) &\equiv \int k(s_1-\omega, |t-\upsilon| \bmod C)\ \xi(\omega,\upsilon)\ d(\omega,\upsilon)
\end{aligned}
$$

where $T(s,t)$ represents the observed temperature at depth d. The posterior densities for the intercept and the slope are plotted in figure 5.5. It is interesting to note that while the trend in temperature is barely noticeable at the surface, the temperature at

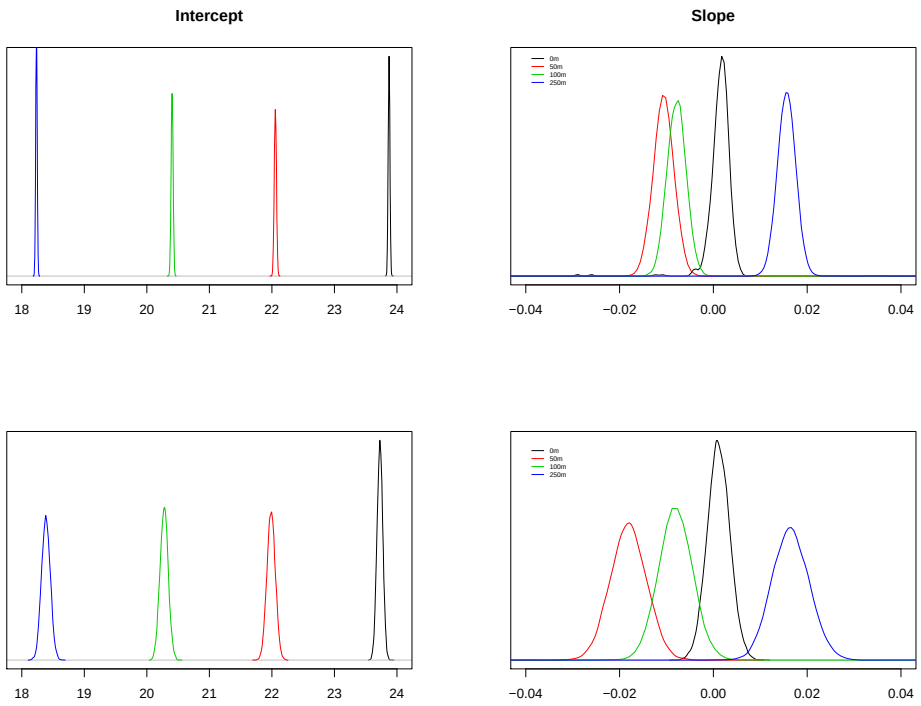


Figure 5.5: Posterior densities for the common mean and annual effect in the subtropical gyre for temperatures at depths $0, 50, 100, 250\ m$. The bottom row includes *cruise effects*.

depths $50m$ and $100m$ are cooling: $.01^oC$ and 0.008^oC per annum, respectively. On the other hand, at $250m$, we observe a significant warming trend of 0.016^oC per annum and observe no trend at $500m$. These results are peculiar, and may be explained if the ocean is absorbing the heat to deeper levels than expected, specifically to $250m$, far below the depth of the Mixed Layer. The cooling at depths $50m$ and $100m$ may then be explained by a new thermocline crossing over the old one somewhere between 100 and 250 meters.

In the second scenario, we investigate annual trends by latitude bands. The design matrix $\boldsymbol{Z}$ now links the depth of the Mixed Layer M to an annual trend for latitude regions $20-24N$, $24-29N$, $29-34N$, and $34-38N$. The posterior distributions of the overall mean and the slopes of the annual trends are plotted in figure 5.6. At first

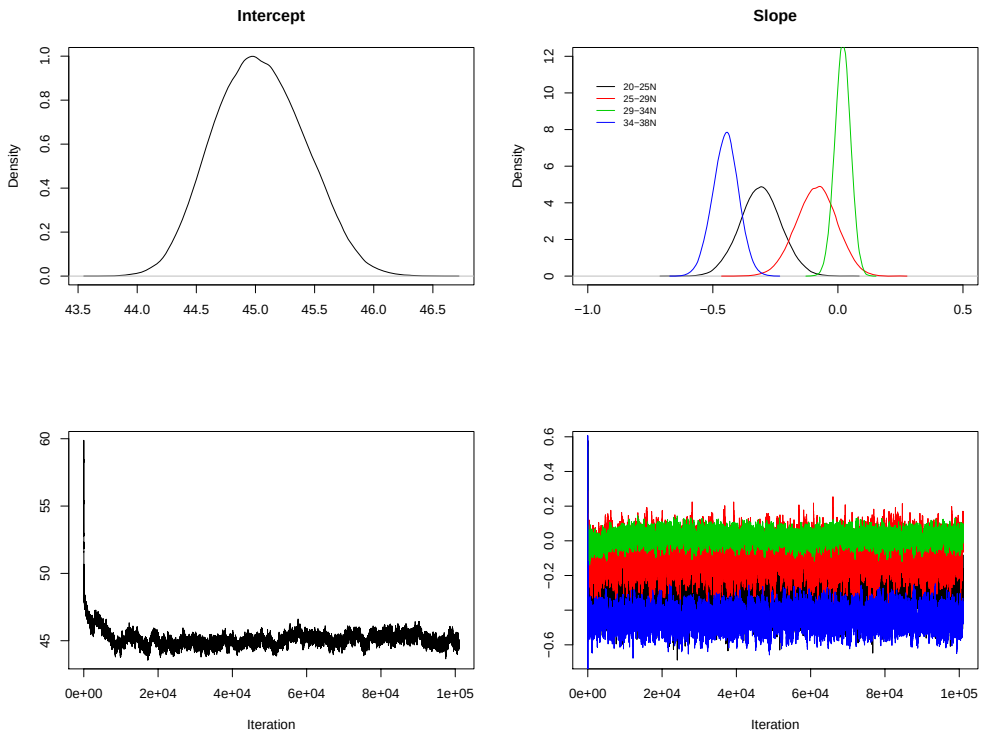


Figure 5.6: Posterior densities for the overall mean and slopes for the trends of the Mixed Layer depth by latitude bands in the subtropical gyre.

glance, it appears awkward that two neighboring regions, $29-34N$ and $34-38N$, are so different. This may be due to the influence of a strong North Atlantic current. We begin to observe the influence around $34N$, which is also the reason we chose this latitude as one of the cut-off points. To better explain changes in the depth of the Mixed Layer, we again look into temperature trends at different depths. Posterior distributions for the overall mean and slopes are plotted in figure 5.7. Interestingly, the most northern region gives the strongest signal of warming at $250m$. This is also the region that is associated with the strongest cooling at the surface and the strongest shoaling of M. In the southern most region, $20-25N$, the entire profile has warmed, most at the surface and least at $250m$. It appears that the increased heat at the surface is transported to deeper waters but at a much slower rate than

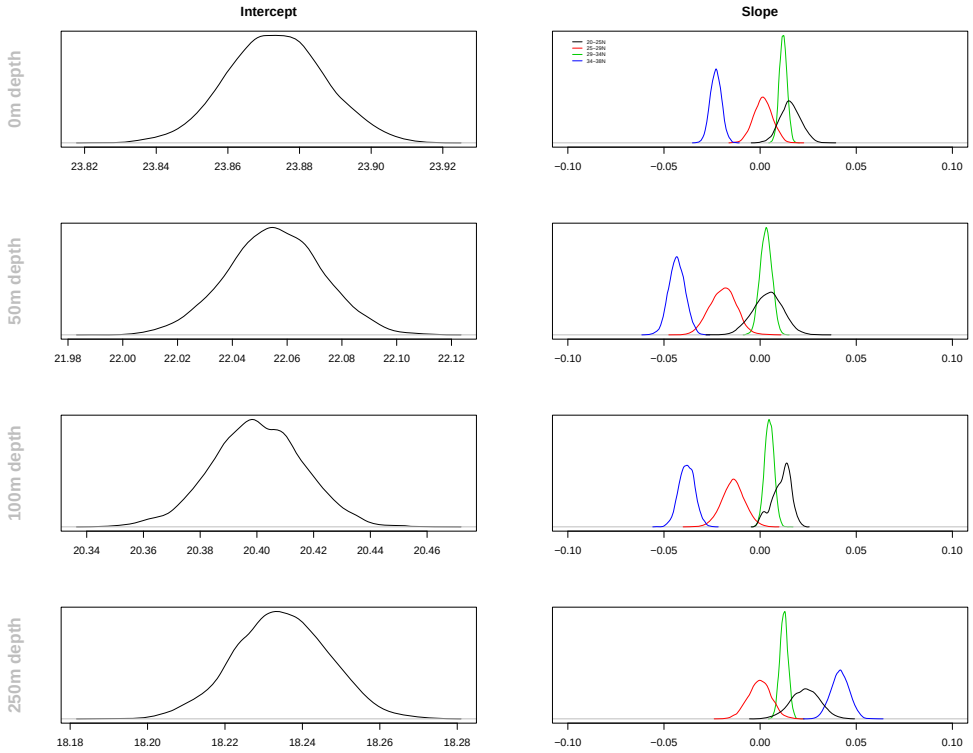


Figure 5.7: Posterior densities for overall mean and slopes for temperature trends at depths $0, 50, 100, 250\ m$ by latitude bands in the subtropical gyre.

in the north. This may be caused by the larger gradient in temperature between the surface and $250m$. The gradient is smaller in the most northern latitude band so the waters mix faster under the increased heat forcing at the surface. After the waters are mixed, the new, shallower Mixed Layer is formed. The shoaling of the Mixed Layer in the southern region is accompanied by significantly less warming at the deeper levels, leading us to conclude that the mixing of waters is probably not complete (see the bottom right panel of figure 5.7).

As a third scenario, we linked M to four different means, corresponding to four different latitude bands and to different decades. The results are listed in table 5.1. The posterior mean for early years, before 1970, is deeper than for more recent decades, as previous analysis suggested. It is also interesting that the latitude bands

Lat. band	$\tilde{\mu}$
34-38N	59m
29-34N	39m
24-29N	45m
20-24N	55m

Decade	$\tilde{\mu}$
<70	57.7m
70's	39.85m
80's	48.4m
90's	48.1m

Counts per decades

Lat.band	<70	70's	80's	90's
34-38N	564	338	254	92
29-34N	582	943	357	579
24-29N	107	894	495	48
20-24N	180	33	166	175

Table 5.1: Posterior mean summaries by latitude band and decade in the subtropical gyre.

which showed the strongest shoaling of M appear to have, on average, a deeper Mixed Layer than the others, suggesting a possible sampling bias. All regions are sampled in relatively similar proportion within a year, with the exception of $34N - 38N$. This region was sampled heavily in the summer months, resulting, perhaps, in an artificially low average value for M.

In conclusion, in the subtropical gyre, although at different rates in different subregions, we are observing a shoaling of the Mixed Layer. One common theme is that larger rates of shoaling of the Mixed Layer are followed with a slight cooling at the shallower depths and strong warming at a depth of 250m. This indicates that, although the mixing of the waters is taking place at different rates, the heat is being absorbed to levels far deeper than M.

5.4.2 Subpolar gyre

There were 2,518 thermal profiles used in the analysis of the subpolar gyre. Interestingly, the Mixed Layer appears to shoal at approximately the same rate as in the subtropical gyre but with greater uncertainty. Introducing *cruise effects* in the subpolar gyre increases the rate of the long term trend as well as the uncertainty about the rate, see figure 5.8. The shoaling of M seems counter intuitive to the oceanographer because scientists have reported the observed cooling of the surface waters in previous studies. Cooling at the surface should deepen the mixing of the waters.

However, this is not what we are observing. In fact, as illustrated in figure 5.9, we see a weak warming trend at the surface and a more significant warming trend below the surface. Therefore, the long term trend of M, in relation to temperature changes, is consistent with that of the subtropical gyre.

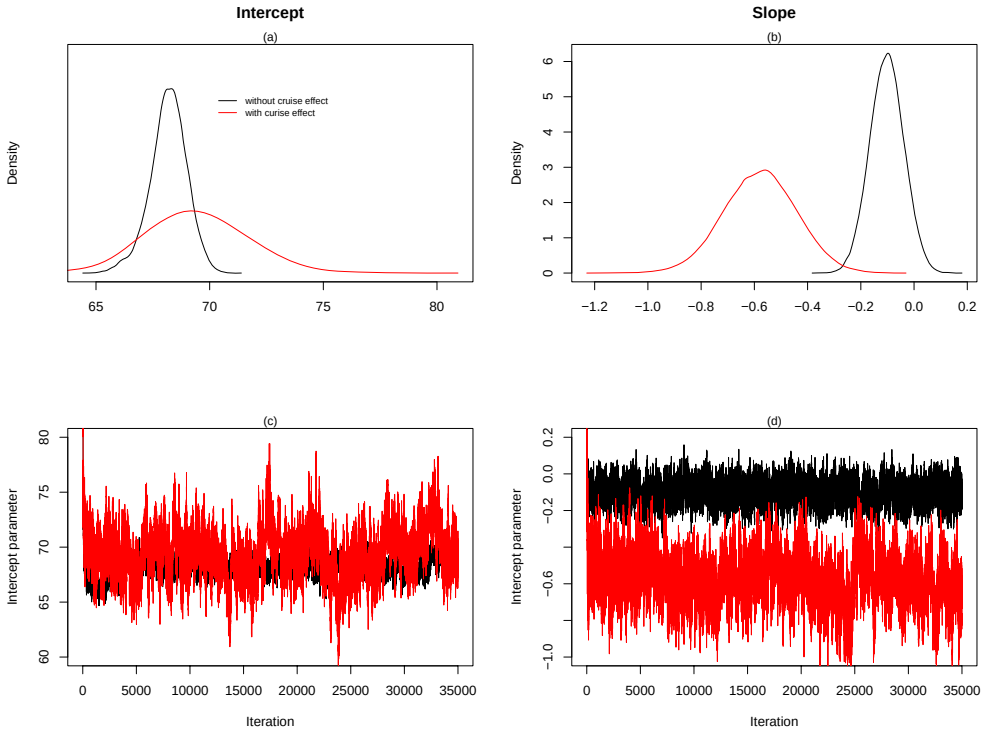


Figure 5.8: Posterior densities for the common mean and the annual effect in the subtropical gyre.

In the subpolar gyre, on average, the posterior means of the cruise effects are larger at higher latitudes and show different trends in different latitude regions (see figure 5.10).

In the subpolar gyre, it is much more complicated to explain the changes in M relative to the changes in temperature. Surface waters in this region move counter-clockwise, and the four latitude bands may be strongly influenced by different sources. Latitude band $45-48N$ is influenced by the strong North Atlantic current, $48-51N$

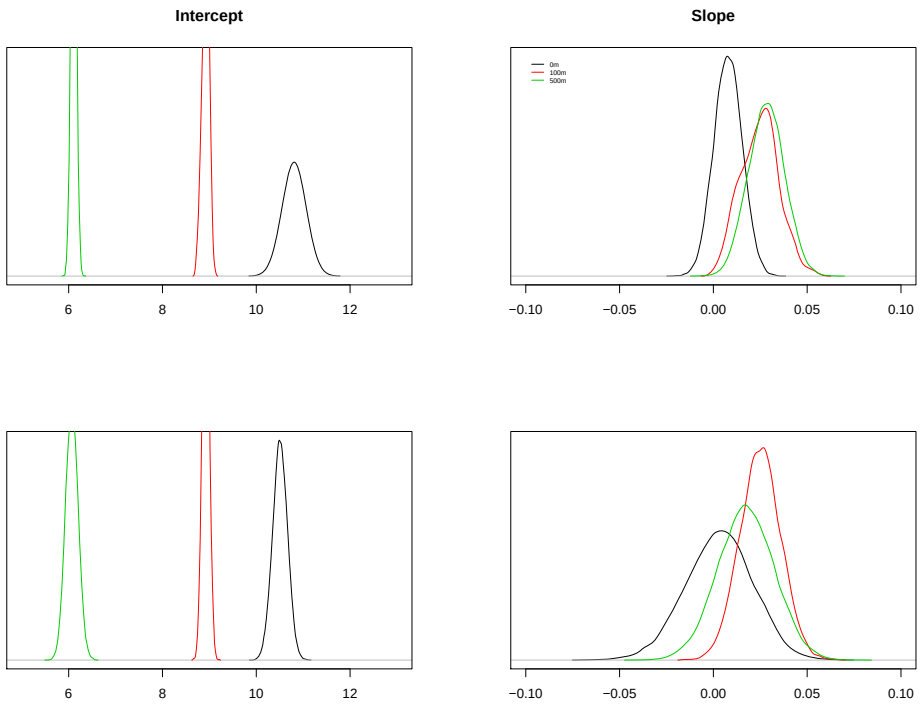


Figure 5.9: Posterior densities for the common mean and the annual effect in the subpolar gyre for temperatures at depths $0, 100, 500m$.

by the cold Labrador current, and $54-58N$ by the cold currents from the North Sea. At the lowest latitude band, $45-48N$, where the strongest surface warming is taking place, we have mixing of the fresh subpolar waters with warmer subtropical waters due to the North Atlantic current. These waters are the least dense so the warming of the profile occurs quickly, but because of the high temperature gradient between the surface and the deep waters, deep mixing takes a long time. Therefore, the new Mixed Layer is shallower than before. The posterior summaries for temperature and M indicate that similar changes are occurring at $51-54N$. Waters have warmed at all depths, and the Mixed Layer has shoaled, leading us to conclude that the waters have also not finished mixing. In the regions of $48N-51N$ and $54-58N$, no significant change in the depth of the Mixed Layer is observed. Additionally, in the region

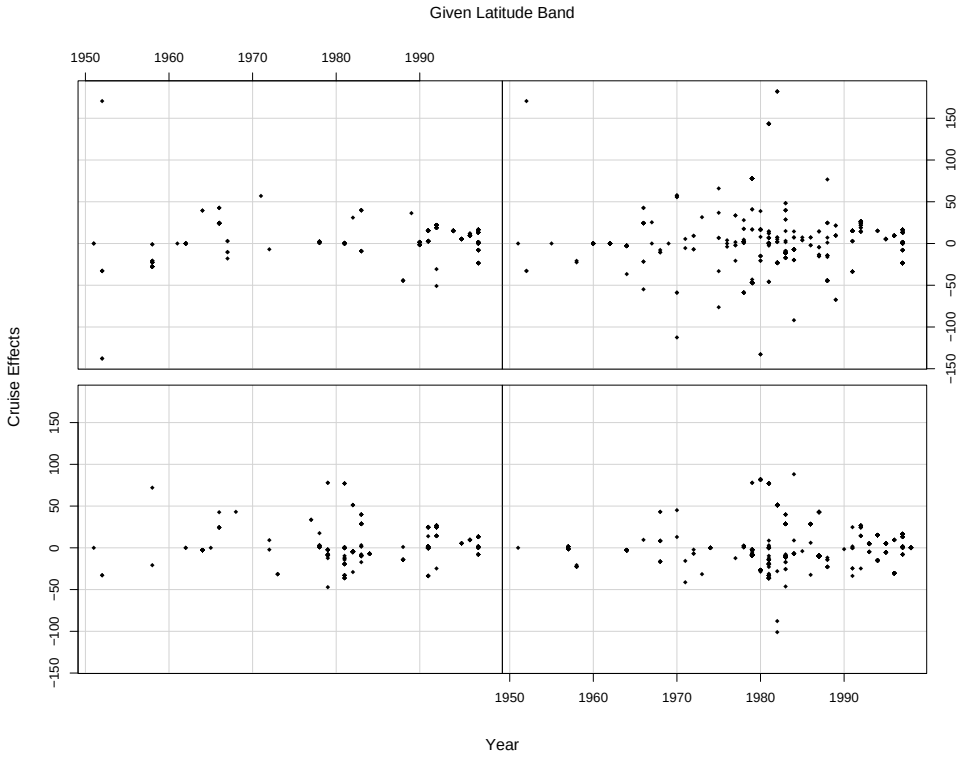


Figure 5.10: Posterior means for the *cruise effects* in the subpolar gyre by latitude bands. Top down by row: 54-58N, 51-54N, 48-51N, 45-48N.

$48N - 51N$, temperatures are notably cooling at all depths. Here, the cold waters of the subpolar gyre meet the warmer, less dense waters brought by the North Atlantic current, which causes the dense waters to sink. Although changes in M, relative to the changes in temperatures, may be interpreted consistently, it is not clear why the trend strengths do not vary smoothly with latitude. One possible explanation is that between $51N$ and $54N$ there exists some very cold, deep currents; this creates a large gradient in temperatures which, in turn, results in slow mixing. By looking at the distribution of temperatures in the subpolar gyre in figure 5.13, this could very well be true. Temperatures in the region $51 - 54N$ have a well pronounced and larger gradient of temperature decline relative to that of the neighboring region, $54 - 58N$.

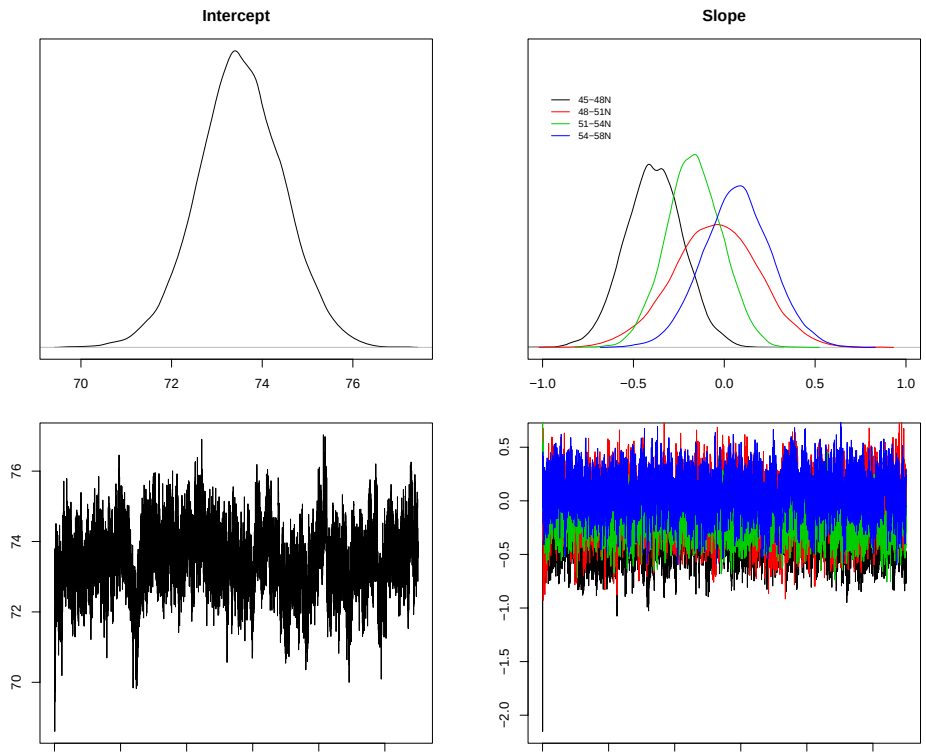


Figure 5.11: Posterior densities for the overall mean and slopes for the trends of Mixed Layer depth by latitude bands in the subpolar gyre.

5.4.3 Goodness of fit

An obvious difficulty that results from not knowing the distribution of the data, $f(x(d)|M)$, is that we cannot assess the goodness of fit using fitted values because there is no one-to-one map between M and the distribution of $x(d)$. We can make predictions and cross validations about M, but we don't have an estimator to compare it with. So to begin assessing the goodness of fit, we plot the likelihood function for each observed profile. In figure 5.5(a)-(j), we plotted an image of the likelihood function for all observed profiles by month and ordered them by year. The likelihood function for each profile is scaled by it's own maximum value, so that the darkest stripe for each profile represents the depths of the maximum likelihood. For better visibility, we added a purple dot at the lowest depth of these intervals and a yellow

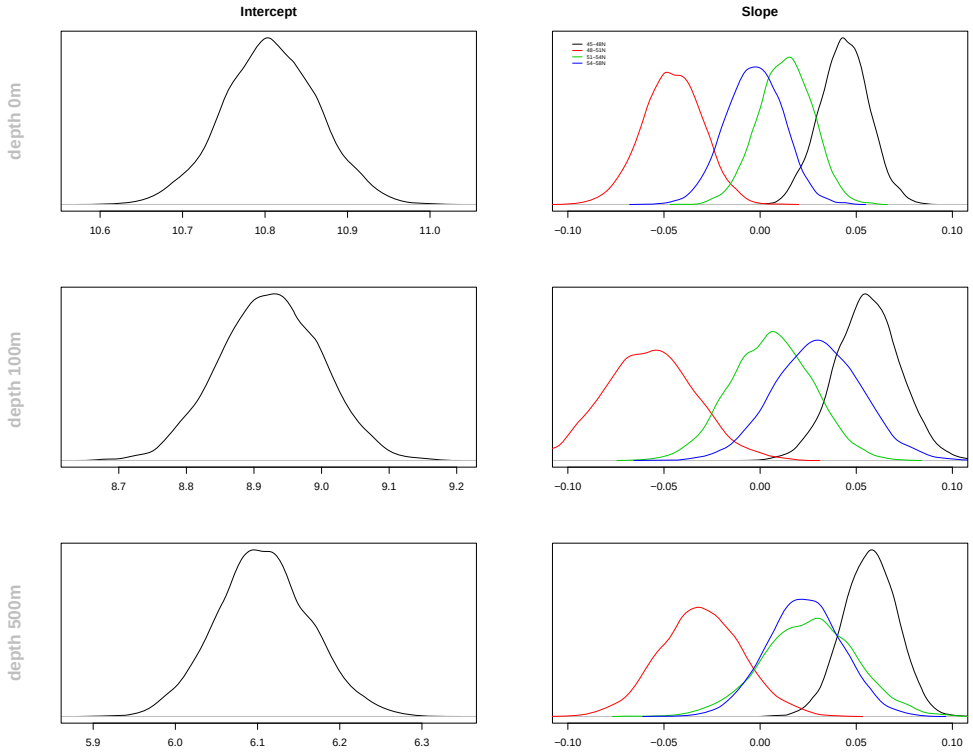


Figure 5.12: Posterior densities for the overall mean and slopes for temperature trends at depths $0, 100, 500\ m$ by latitude bands in the subpolar gyre.

dot at the lowest depth of the intervals with the second largest likelihood. The blue, red, and green dots in figure 5.5 denote the posterior mean, approximate median, and the approximate mode, respectively. Overall, we see that the posterior mean has adapted well to the likelihood function. The largest discrepancies are found in the months of March and April; the months where rapid changes may occur. In $< 2.9\%$ of the observations taken between January and April, the maximum likelihood estimate of M is larger than $300m$, while the posterior mean never dips below $250m$. Although a relatively small number of observations fit this criteria, they potentially have a large influence on the posterior mean and the estimates of other parameters. To understand whether or not these observations are outliers, we need to understand how much information is contained in the maximum likelihood value

Lat. band	$\tilde{\mu}$
54-58N	78.7m
51-54N	101m
48-51N	70.7m
45-48N	58.6m

Decade	$\tilde{\mu}$
<70	88.7m
70's	92.7m
80's	55.6m
90's	55.6m

Counts per decades

Lat.band	<70	70's	80's	90's
54-58N	85	11	30	240
51-54N	105	216	207	171
48-51N	40	53	143	148
45-48N	92	144	547	286

Table 5.2: Posterior mean summaries by latitude band and decade in the subpolar gyre.

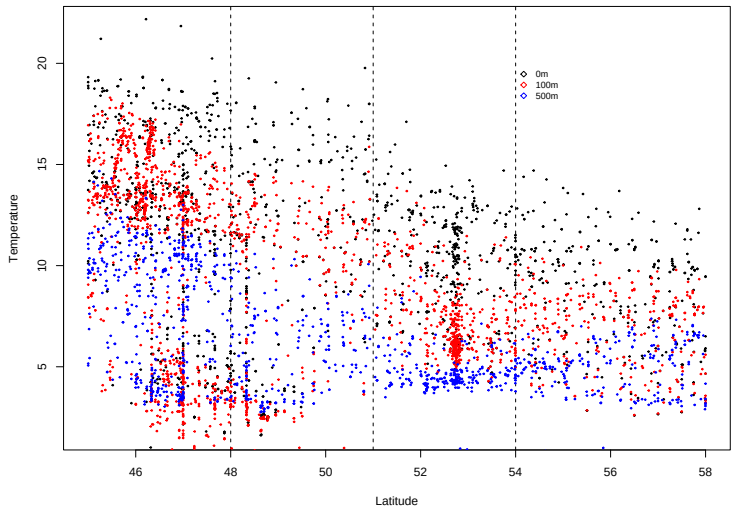


Figure 5.13: Distribution of temperatures with latitude in the subpolar gyre.

estimator, and we need to understand the information of the neighboring profiles in space. The likelihood function may have one, two, or more depth intervals with a high likelihood. Moreover, the high likelihood regions may be close to or far from each other. To address these issues, we evaluated the likelihood function at sampled values of M, and then we compared it to the distance between the posterior mean ($\tilde{\mu}$) and the minimum depth of the maximum likelihood interval ($\hat{\mu}$) in figure 5.10. The likelihood values have, again, been normalized by their own maximum so that the likelihood value is between zero and one for each profile. The size of the points in

figure 5.10 are proportional to the conditioning variable listed in the legend in each of the six panels. As a conditioning variable, we used :

1. the distance between $\hat{\mu}$ and the depth of the second largest likelihood value ($\hat{\mu}_2$),

2. the distance between $\tilde{\mu}$ and the depth of the second largest likelihood value ($\hat{\mu}_2$),

3. the value of the second largest likelihood ($lik(\hat{\mu}_2)$) relative to the maximum likelihood ($lik(\hat{\mu})$),

4. the average posterior density of the mean process μ,

5. the distance between the posterior mean ($\tilde{\mu}$) and the weighted mean of the neighboring $\hat{\mu}$'s including the observations from the current expedition, and

6. the distance between the maximum likelihood estimate ($\hat{\mu}$) and the weighted mean of the neighboring $\hat{\mu}$'s excluding the observations from the current expedition.

The legends in figure 5.10 correspond to the $\leq 50^{th}, 75^{th}, 85^{th}, 95^{th}$, and 99^{th} percentile of the conditioning variable. The points with the largest *residual values*, *i.e.* $|\hat{\mu} - \tilde{\mu}| \geq 150m$ and the average likelihood greater than 0.3, have a large distance between the two likelihood modes, while those with a large *residual value* and low average likelihood have a relatively moderate distance between the two modes, see the top panel of figure 5.10. From the second panel of 5.10, we see that the observations with large residuals and a high average likelihood value are those where there are at least two distinct modes in the likelihood, and the posterior mean $\tilde{\mu}$ favors the second, shallower mode. All profiles that have a large residual and favor the second largest likelihood value are the profiles where there are two distinct $\hat{M}'s$. The distance

between the depths with the two largest likelihood values need not correspond exactly to the two largest likelihood modes, but it is a good approximation. The large residuals with a low likelihood correspond to the observations where the posterior mean is closer to $\hat{M}$. A low value of the likelihood may indicate influence of the neighbors or over-smoothing. The value of the second largest likelihood, relative to the maximum likelihood, is reflected as a conditioning variable in the middle panel. The largest 15% of $lik(\hat{M}_2)$ are highlighted in blue.

The first three panels give us insight to the shape of the likelihood function for observations with large residuals. It is also important to understand the information given by the neighboring observations. We need to be reassured that there are no single profiles with an artificially high influence on the mean. For example, the likelihood function is made to fit the oceanographer's belief for the depths with a large probability. In the regions of low probability, where we don't fully trust the assessments, we may have a large likelihood ratio between the two unlikely values of M. As previously noted, the likelihood function was bounded away from zero and set to a constant to avoid such problems. Nevertheless, we must persuade the reader that no such observations occur.

First, we note that all of the observations with large residuals have a moderate to high value of the average posterior density of the mean. In other words, the observations with large residuals are unlike their neighbors, and hence, the model favors the mean process rather than the likelihood. In the fourth panel of figure 5.10, we are conditioning on the average posterior density of the mean process. The largest 15% are highlighted in blue. It is reassuring to know that the highest density of the mean process is in agreement with the high likelihood values.

To assess whether or not these large residuals correspond to the outliers, we created a new variable: the weighted mean of the neighboring maximum likelihood

within a range of seven days and two degrees of latitude. As weights, we used a correlation function dnorm(d,0,30)/ dnorm(0,0,30) in time (units of days) and dnorm(d,0,5)/ dnorm(0,0, 5) in space (in units of degrees latitude). In the fifth panel, the weighted mean was computed including the observations taken on the same expedition but not the observation itself, and in the sixth panel, the weighted mean was computed excluding the observations taken on the same expedition. In the fifth panel of 5.10, we used the difference between the weighted mean and the posterior mean as a conditioning variable, and in the bottom panel of 5.10, we used the difference between the MLE and the weighted mean . The weighted mean is larger than the posterior mean for all observations with large residuals. However, the weighted mean for these observations are still much closer to the posterior mean than to the MLE estimator, indicating possible over-smoothing. In the bottom panel, we notice that those profiles which have the largest residuals are also very different than the weighted average of their neighbors, explaining why the model prefers the neighboring observations. The results for other data sets and models are listed in Appendix A.

To assure ourselves that there are not many observations that have a high average likelihood value and a large posterior density of the mean process, but large residuals, we searched for the observations that fit in the top 20% of all three criteria. These would be the observations that have a strong influence on the posterior mean. There were six such observations, all of which had a clearly defined Mixed Layer depth, and therefore, we have a strong belief about M.

5.5 Conclusion

The goal of this study was to assess the changes in the depth of the Mixed Layer in Atlantic Ocean. We gave a novel approach to weighing the information provided

by the profiles to determine the uncertainty about M. This is the first probabilistic approach to such a task. Additionally, we provided an example of modeling spatio-temporal dependence among the observed profiles. Our model can accommodate many ways to explore the long term changes in M. For example, we analyzed linear trends, trends by latitude regions, mean levels by decades, and mean levels by latitude regions, but the model is easily extended to other choices. Our analysis points toward the belief that, in both the subtropical and subpolar gyre of the North Atlantic Ocean, the Mixed Layer has been shoaling. In both regions, we notice significant warming trends of the deeper waters, indicating that the ocean is storing heat at deep levels. Temperature trends at deeper levels are also better than surface temperatures at indicating trends in M. Shoaling of the Mixed Layer in the subpolar gyre is contradictory to the previous belief. However, the methods used in those analyses are quite different from ours. In the near future, it will be possible to compare the methods on the exact same data sets which will give us the final reassurance about the long term depth of the Mixed Layer.

(a) January

(b) February

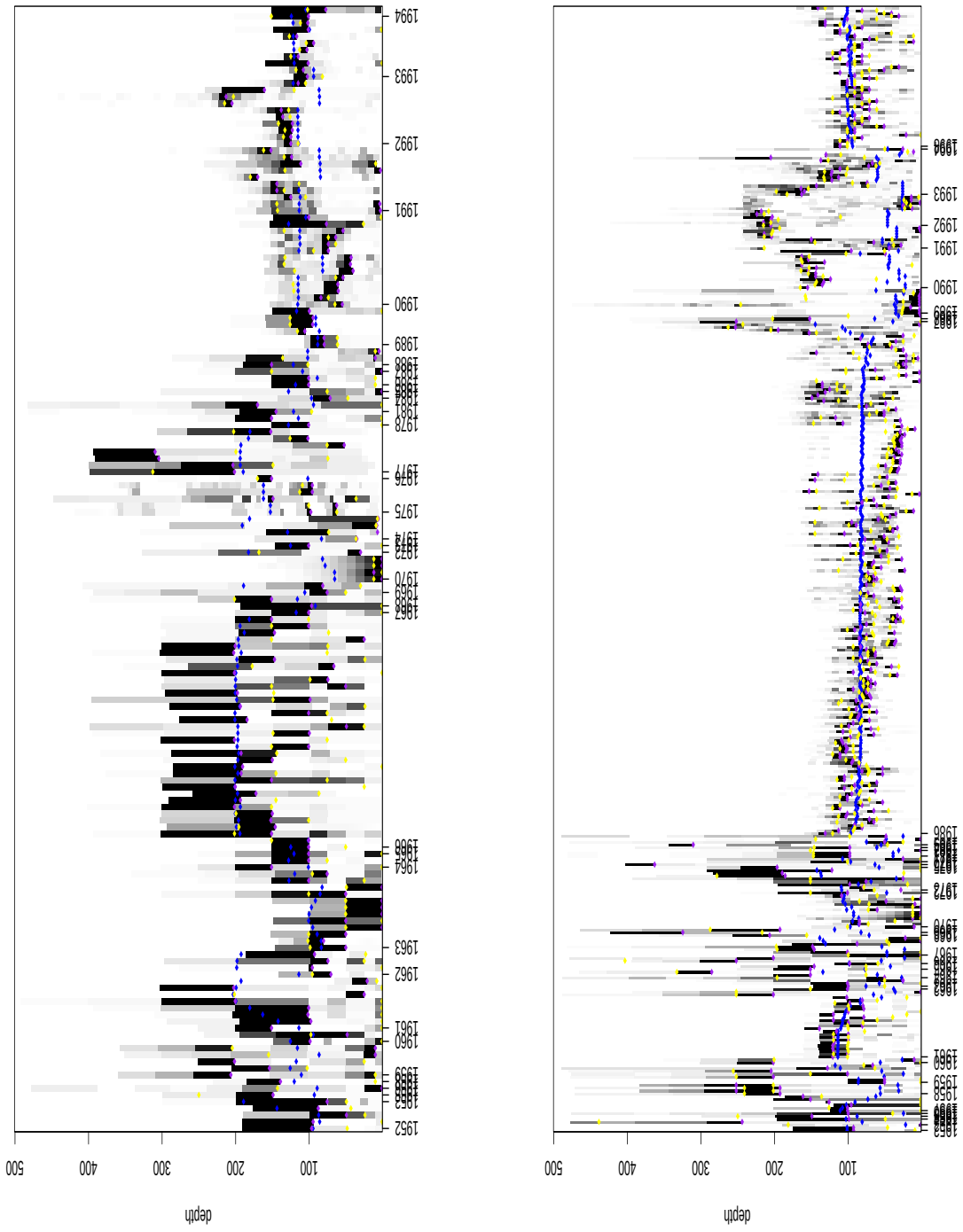
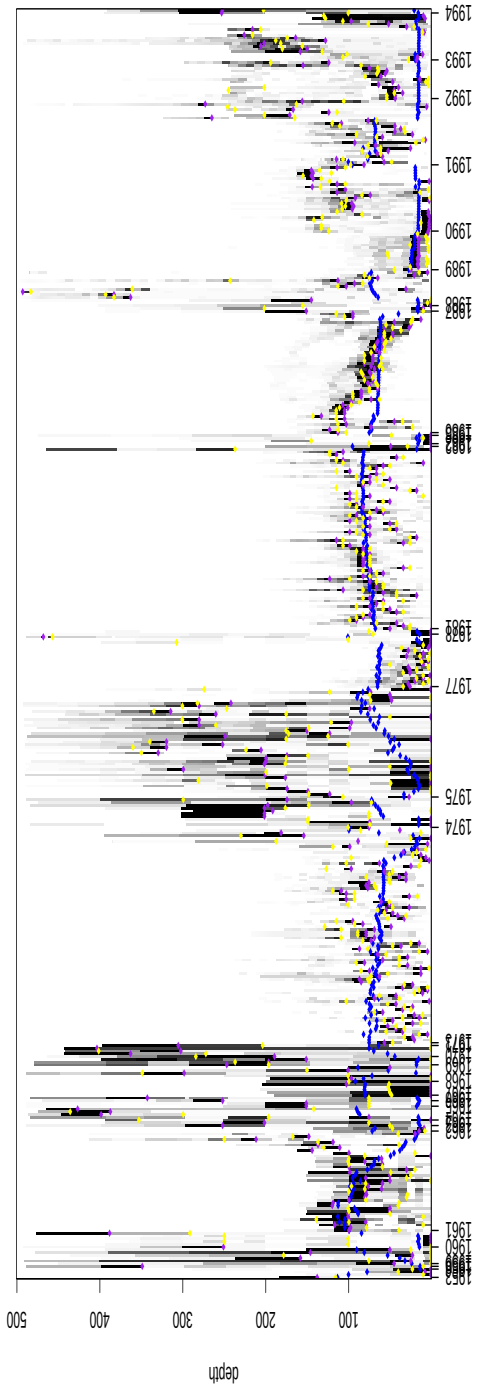
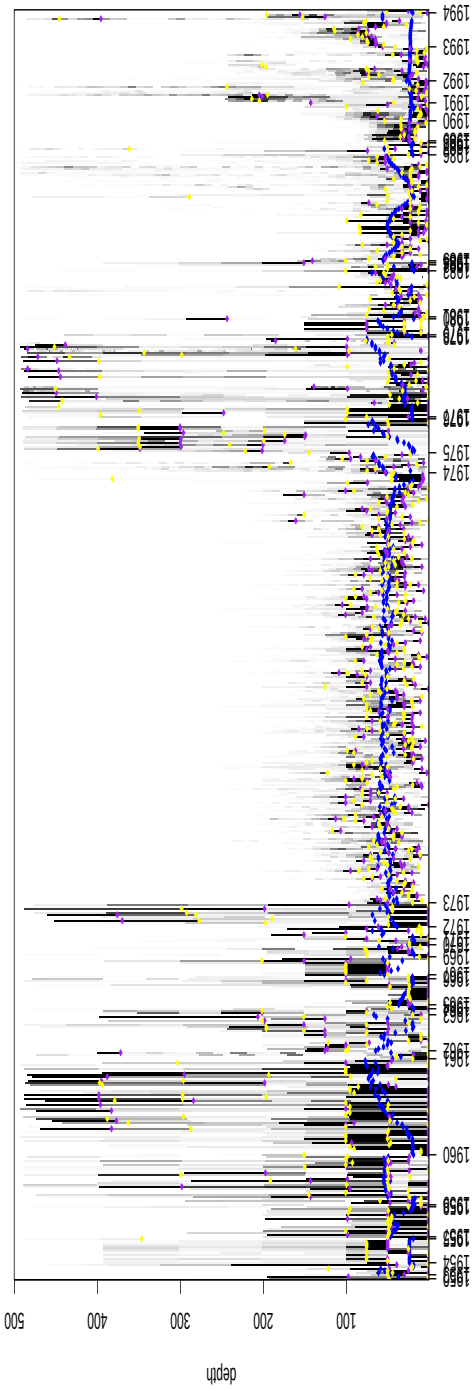


Figure 5.14: Likelihood function and the posterior mean.

(c) March



(d) April



(e) May

depth
100
200
300
400
500
1960
1973
1974
1980
1982
1983
1986
1990
1992
1993
1994

(f) June

depth