Bayesian approach of pollen-based palaeoclimate reconstructions: Toward the modelling of ecological processes

Vincent Garreta

▶ **To cite this version:**

Vincent Garreta. Bayesian approach of pollen-based palaeoclimate reconstructions: Toward the modelling of ecological processes. Climatology. Université Paul Cézanne - Aix-Marseille III, 2010. English. NNT : . tel-00495890

HAL Id: tel-00495890

https://theses.hal.science/tel-00495890

Submitted on 29 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Paul Cézanne Aix-Marseille III

N° 2010AIX30010

Approche bayésienne de la reconstruction des paléoclimats à partir du pollen : Vers la modélisation des mécanismes écologiques

Thèse
pour obtenir le grade de

Docteur de l'Université Paul Cézanne
Faculté des Sciences et Techniques en

Géosciences de l'environnement

Vincent Garreta

Dirigé par

Dr. Joël Guiot et

Dr. Christelle Hély-Alleaume

Ecole doctorale Sciences de l'environnement (ED251 - EDSE)

Jury

Pr. Edouard Bard	examinateur	(Collège de France - CEREGE)
Dr. Pascal Monestiez	examinateur	(INRA - Unité BioSP)
Dr. Eric Parent	rapporteur	(ENGREF - AgroParisTech)
Pr. John W. Williams	rapporteur	(Univ. Wisconsin-Madison - Dpt. of Geography)

Avril 2010

Résumé

Le pollen conservé dans les sédiments lacustres constitue un indicateur essentiel pour reconstruire l'évolution de la végétation et du climat passés sur les continents. Actuellement, les reconstructions climatiques se basent sur des modèles statistiques décrivant le lien climat-pollen. Ces modèles posent des problèmes méthodologiques car ils sont tous basés sur l'hypothèse que la relation pollen-climat est constante au cours du temps, impliquant que les paramètres non climatiques déterminant cette relation aient une influence faible. Cela est contredit par les développements récents en écologie et en écophysiologie. C'est pourquoi, dans ce travail, nous développons une approche intégrant un modèle dynamique de végétation et les processus majeurs liant la végétation au pollen capté par les lacs. Le cadre bayésien fournit une base théorique ainsi que les outils pour inférer les paramètres des modèles et le climat passé. Nous utilisons ces nouveaux modèles pour reconstruire le climat de l'Holocène en différents sites européens. Cette approche qui permettra des reconstructions spatio-temporelles requiert encore des développements autour de l'inférence de modèles semi-mécanistes.

Géosciences de l'environnement

Mots-clés Paléoclimat, paléo-végétation, fonction de transfert, modélisation hiérarchique bayésienne, modèle dynamique de végétation, pollen, DGVM, dispersion pollinique, Europe

Réalisé au sein de l'équipe "Ecosystèmes continentaux et marins", CEREGE UMR 6635 (Unité mixte CNRS, Université Paul Cézanne, IRD, Collège de France), Europôle Méditerranéen de l'Arbois, 13545, Aix en Provence, France

Bayesian approach of pollen-based palaeoclimate reconstructions: Toward the modelling of ecological processes

Abstract

Pollen preserved in lacustrine sediments form a crucial archive for reconstructing past climate and terrestrial vegetation change. Currently, climate reconstructions are based on statistical models describing the link between climate and pollen. These models raise methodological problems because they are all based on the hypothesis that climate-pollen relationships are constant over times; implying that non-climatic parameters driving the relation have weak influence. This is not in agreement with recent developments in ecology and ecophysiology. That is why, in this work, we develop an approach integrating a dynamical vegetation model and major processes linking vegetation and pollen trapped in lakes. The Bayesian framework provides us with a theoretical basis and tools for the inference of model parameters and past climate. We use these new models for reconstructing Holocene climate in various European sites. This approach, which may allow spatio-temporal reconstructions still requires developments around statistical inference for semi-mechanistic models.

Environmental Geosciences

Key words Palaeoclimate, Palaeo-vegetation, Transfer Function, Hierarchical Bayesian Modelling, Dynamical Vegetation Model, pollen, DGVM, pollen dispersal, Europe

Contents

Introduction

Climate is an Earth-scale system made of interrelated components - the atmosphere, oceans, continents and cryosphere. These components interact, for example, by exchanging energy, water and carbon at different time scales. The global functioning and exchange rates of this system are driven by external and internal influences. These influences, called 'forcings', include the variations in insolation, solar and volcanic activities and greenhouse gases (e.g. Milankovitch, 1941; Berger, 1977; Solomon et al., 2007). Each of these forcings varies at different time scales, sometimes largely exceeding the length of instrumental climate records (a few hundreds of years or so). For example, insolation varies according to the Milankovitch cycles, with characteristic phases of 100, 41 and 21 thousands of years (e.g. Berger and Loutre, 2004).

The understanding of climate dynamics is achieved in part through the study of the various responses to these forcings. Such a task requires observations on time series far longer than instrumental records. Time series of past climatic conditions are often reconstructed from remains of ancient living organisms acting as climate surrogates called 'proxies'. For example, common proxies of the past climate on continents could be pollens and diatoms preserved in lake sediments. These reconstructions provide benchmarks for testing climate models on a large range of states and forcings having no analogue in the instrumental period (e.g. the works related to projects COHMAP, PMIP and PMIP2, COHMAP Members, 1988; PMIP Participants, 2000; Braconnot et al., 2007).

However, obtaining such reconstructions remains a multidisciplinary challenge. The proxies collected, measured and dated from sediment cores provide an incomplete information about past biotas, which themselves have complex relations with climate. The understanding of the relations linking the environment (including climate), the biota and the proxy recorded in sediments is fundamental to a proper climate reconstruction. Mathematical modelling of these links is necessary to provide quantitative estimates of the past environmental conditions.

Pollen as a climate record

In this thesis, we developed methods for reconstructing past climate recorded by pollen assemblages. Such sporopollinic data is retrieved from cores extracted from lake sediments. A sediment sample is sieved and chemically cleaned from its carbonate and silicate components (Faegri and Iversen, 1964) and pollen grains are identified based on modern reference collections. Identification based on the morphology of pollen grains is possible at the species, family or genus level depending on the pollen type. Then, pollen assemblages consist of counts of pollen grains per taxa, relatively to a total number of counts in the sample. Dating of a sample sequence along the core is performed using radiocarbon measurement. Each pollen sample is associated to a date and an uncertainty on this date. Uncertainties arise from radiocarbon measurement errors, the transformation into calendar ages and the extrapolation of the dated levels to the whole core. Pollen assemblages are noisy records of the vegetation surrounding the lake filtered through a chain of processes including the pollen production by plants, its dispersion by winds, its capture and concentration in the lake and its preservation in sediments (Prentice, 1985).

In turn, vegetation species record their environment through their absence/presences and productivity. The environmental conditions having a potential impact on plant species can be divided in abiotic (or 'physical') factors and biotic (or 'biological') factors. Abiotic factors include mean and extremes of climatic variables (e.g. temperature, precipitation, cloudiness), but also CO_2 concentration, soil type and nutrient availability. A biotic factor condition are for example, the presence of competitors for lights and/or water and predators (e.g. Woodward, 1987; Bugmann, 2001).

The number and the complexity of the processes linking climate and pollen composition of sediments complicate the direct interpretation of pollen assemblages in terms of climate variables. Qualitative investigations of such records, often in terms of presence/absence of taxa, give indications about vegetation type and possibly, climate type. Quantitative vegetation and climate reconstructions are achieved trough the building of mathematical models.

Classical approaches for pollen-based reconstructions

In classic pollen-based palaeoclimate reconstructions, the climate-pollen relation is considered as a whole. In other words, there is no explicit modelling of climate-vegetation processes nor how pollen is produced by vegetation. Such models are purely statistical and called Transfer Functions (TF, e.g. Brewer et al., 2007). Until the early works of Iversen (1944), TF were correlative, i.e. they describe the response of one variable (climate or pollen) to the variations of the other one without accounting for functional relations. These models are calibrated on modern pollen and climate distributions, and have been used to reconstruct climate over the last tens or hundred of thousand years (e.g. Guiot et al., 1989, for a 140,000 years reconstruction and a review of classical TF in chapter 1).

In using such relations for palaeoclimate reconstructions, one assumes that (i) climate is the main driver of vegetation changes, (ii) pollen-climate relations are constant over time and (iii) the vegetation is in steady equilibrium with climate. These models are useful and were the only solution in times when we had poor knowledge about the vegetation functioning, a reduced number of statistical models, and limited computing capacities. Their application may appear today as a 'brute force' use of the uniformitarianism principle: 'The present is the key to the past' (Hutton, 1795; Alley, 2001). Indeed, the lack of causality and process modelling in the TF may lead to an overly simplistic interpretation of present climate-pollen distributions for being used in an environment that is known to be significantly different from the present. An example is atmospheric CO_2 concentration, which cannot be included in the TF since its spatial variation is negligible. The problem, raised by Cowling and Sykes (1999), is that CO_2 is expected to have a significant effect on plants, changing their optimal climatic range and their resistance to drought. Since, in the past, CO_2 concentration varied in a range of values (between 170 and 300 ppm over the last 450 thousands of years) lower than present (exceeds 380 ppm), reconstructions obtained from classical TF are expected to be biased.

Mechanistic approaches in pollen modelling

A solution for reconstructing climate in a better agreement with our recent knowledge of the relations linking climate and pollen is to explicitly model the linking processes. The first step in this direction has been made by Guiot et al. (2000), proposing to insert a mechanistic vegetation model in the TF. This vegetation model (BIOME3, Haxeltine and Prentice, 1996) was a process-based model representing the photosynthesis, nitrogen allocation and accounting for CO_2 changes. Guiot and coauthors used a statistical 'matching' between model outputs (the vegetation) and pollen to connect them. They proposed a Bayesian statistical model (Robert, 2001; Gelman et al., 2004) and its associated Markov Chain Monte Carlo algorithm (Robert and Casella, 1999) for reconstruction, i.e. the inversion of the computer model based on pollen to obtain climate.

In parallel to the development of TF, generations of models have been designed to relate the vegetation to its remains recorded as pollen assemblages in sediments. Until the 'R-value' model (Davis, 1963) they were based on the processes expected to link the vegetation to its pollen record, including pollen production, dispersion and accumulation. Early models and studies were based on a few mechanistic relations, mainly the dispersion modelling using Sutton's equation (Tauber, 1965; Webb, 1974; Parsons and Prentice, 1981; Prentice, 1985). They have evolved into approaches merging mechanistic and statistical modelling and using simulations and real data for validations (e.g. Sugita, 2007a,b, and references therein) or full Bayesian modelling and inference (Paciorek and McLachlan, 2009).

These works provide further opportunities for TF modelling at two levels; (i) by providing process-based structures for the pollen-vegetation links, and (ii) by demonstrating the usefulness and accessibility of mechanistic/statistical coupling in a Bayesian framework.

Manuscript content

In this manuscript we develop a fully process-based TF. The final TF includes a Dynamic Vegetation Model (DVM, LPJ-GUESS, Smith et al., 2001) for describing the dynamic links between climate and vegetation, which is coupled to a statistical model of pollen-vegetation relationships that is based on pollen production, dispersion and accumulation processes. This composite structure formed with a mechanistic climate-vegetation model and a statistical vegetation-pollen model is framed in a statistically sound framework - Bayesian hierarchical modelling - that allows a proper characterisation of the calibration and reconstruction exercises and provides us with various algorithmic tools for inference.

The manuscript chapters are under the form of articles (second chapter is published in *Climate Dynamics* and third chapter is in review for *Environmetrics*). In the first chapter, we review the existing TF, classify them and describe their main hypotheses. We show the similarity between a class of TF and the Species Distribution Models (SDM) used in Ecology. We propose a framework to exchange tools and ideas between both disciplines. In the second chapter, we present a Sequential Monte Carlo algorithm (SMC) allowing the inversion of the dynamic vegetation model LPJ-GUESS for climate reconstruction. The methods is applied to a high-resolution sediment core covering the Holocene in Southwest Germany. In the third chapter, we present the building and calibration of a process-based model to link vegetation simulated from LPJ-GUESS and pollen. This model includes pollen production, dispersal, and accumulation for continental scale datasets as used in pollen-based palaeoclimatology. In the last chapter, we merged both approaches (dynamic inversion and spatial calibration) and use them for climate reconstruction at four Swedish sites. Emphasis is placed on the challenge of the inference of statistical models including mechanistic components.

Chapter 1

Correlative versus process-based niche models in palaeoclimatology and their relation with ecology

Abstract Palaeoclimate reconstructions are based on mathematical models of the relation between the habitat (including climate) of the organism and its remains measured in sediments. These models are called transfer functions (TF). The majority of existing TF are correlative, i.e. they statistically describe the organism response to climate (or the inverse) without accounting for processes expected to drive the response. The correlative way of modelling implies to assume an irreducible set of hypotheses: a perfect 'dynamical equilibrium' of the vegetation with regard to climate and a negligible change in the vegetation response under different CO_2 concentrations. These hypotheses are either in contradiction with actual knowledge in ecophysiology or not robustly testable. This has a major impact on the reliability of the reconstructions, especially outside the modern range of environmental conditions. A process-based approach developed around the a dynamical vegetation model provides a necessary complement for more accurate palaeoclimate reconstructions.

Based on a contemporary vision of the ecological niche theory, we review the hy-

potheses of correlative models used in palaeoclimatology. We compare them to the species distribution models (SDM) used in ecology for predicting future climate change impact. In this field, the same debate and the same lack of process-based approaches raises questions about the predictive power of SDM.

Thus, the reconstruction of past climate from pollen and the prediction of future plant species distributions face the same challenge. We describe a Bayesian framework in which the core tools - the (niche) model for the environment-plant relations - are exchangeable between disciplines. We believe that this framework will encourage interdisciplinary cross-fertilisation and allow the rapid development of improved process-based models.

1.1 Introduction

In palaeoclimatology, the models representing pollen-climate relationships, referred to as Transfer Functions (TF), are models of the species response to a finite number of environmental variables. They are therefore based on the hypothesis that species have intrinsic limits defining their 'niche' through complex processes (e.g. competition for resources, physiological limits), and whose role is prevalent over other environmental factors (see e.g. Huntley, 1996; Jackson and Overpeck, 2000, for a discussion of these hypotheses in palaeoclimatology).

TF can be classified in two distinct types. Type I TF are formed by the correlative models lying on the statistical mapping of the species response to its environmental or its inverse. Type II TF are formed by process-based models, i.e. models whose structure is based on the processes expected to create the species responses to their environment. As reviewed in this article, type I methods are subject to two, hardly testable, hypotheses: 'dynamic equilibrium' of vegetation with respect to climate (Webb III, 1986; Prentice, 1986) and that the distribution of the plant is robustly controlled by a small set of climate variables (discussed in early TF article cited hereafter). Moreover, calibrated on a - finite - modern range of environmental conditions, it is well-known that non-parametric TF (those defined locally by data) cannot be used outside this range (know as the 'no-analogue' problem, see e.g. Williams and Jackson, 2007). When parametric TF (defined everywhere by a parametric structure) are used outside the modern range, this add another hypothesis, often supported by theoretical arguments but not testable using modern datasets (e.g. Gaussian and symmetric structures in Vasko et al., 2000; Gonzales et al., 2009).

The under-development of type II TF raises questions about the reliability of reconstructions, which are all based on the same two hypotheses. For instance, despite their diversity in shape (presented hereafter), correlative TF depend on a common set of approximations making their reconstructions correlated (e.g. exclusion of the same environmental variables from their niche: CO_2, soil type and absence of modelling of present or past vegetation dynamics). For a set of reconstructions produced from these

TF, (i) an agreement between methods is a weak indication of reconstructions accuracy, at least outside the range of calibration, and (ii) a divergence between methods cannot be corrected, for example, by selecting the intersection of the various reconstructions. The solution is to be found in developing models based on different assumptions, as credible as those included in the correlative TF. The process-based TF provides this independent complement.

The correlative (type I) models describe the link between pollen and its environment (usually climate) through purely descriptive relationships. They date from the beginning of quantitative palaeoclimatology and are sufficiently numerous to be separated in three groups based on their mathematical form (see Figure 1.1). In the following, we briefly classify the various methods used.

The type I.1 TF includes models of the climate distribution as a function of pollen assemblages. We call them 'backward' TF because their modelling is inverse to the causative relation 'climate drives species'. They include TF relating climate distribution to the taxa's presence: Indicator Species (IS, Iversen, 1944), whose updated version are the Mutual Climatic Ranges method (MCR, Atkinson et al., 1986) and the Probability Density Function method (PDF, Kühl et al., 2002). They also include models relating climate to quantitative indicators of the species, e.g. linear model in Bartlein et al. (1984), Artificial Neural Networks (ANN) in Peyron et al. (1998) and Generalised Additive Model (GAM) in Gersonde et al. (2005).

The Modern Analogue Technique (MAT, Hutson, 1980; Overpeck et al., 1985; Guiot, 1990) is often separated from the other methods (Jackson and Williams, 2004) and said structure-free because it does not explicitly require to make assumptions about the shape of the pollen-climate relation. In this method, climate corresponding to a pollen assemblage y is reconstructed as the climate corresponding to pollen assemblages from the modern dataset that are analogous to y. We propose to interpret it as a quantitative type I.1 TF, in which climate is locally smoothed in a pollen space.

The type I.2 is said 'direct' because it includes TF describing the pollen assemblage as

a function of climate. In palaeoclimatology, these models are called 'response surfaces' and interpreted as maps of the pollen response to climates. They are regression models with different shapes (polynomial, Gaussian curves, non-parametric smooth response) and include the Response Surface (RS, Bartlein et al., 1986), the Bayesian Multinomial Gaussian Response (BUMMER, Vasko et al., 2000) and its updates (e.g. Bhattacharya, 2006), the Bayesian semi-parametric response surfaces (Haslett et al., 2006), and the Expanded Response Surfaces (ERS, Gonzales et al., 2009).

The type I.3 TF is said to be 'not directed' and it consists of the Weighted Average-Partial Least Square (WA-PLS) method (ter Braak et al., 1993) and its variants. These methods are based on a PLS regression projecting both environmental variables and pollen assemblages to a new space made of latent components. In this sense, it does not consider any direction in the relation between pollen and environmental variables.

The type II, composed with process-based models, is not common in the literature and contains only mechanistic relations (review in Guiot et al. (2009) and classification figure 1.2). 'Process-based' means that core equations describe the pollen-plant-environment system through its constituent physical, biogeochemical and competition processes. By including such equations summarising the current knowledge on functioning, a process-based approaches is more likely to represent a credible climate-plant link - outside - the range of modern climate. This type of modelling was initiated by Guiot et al. (2000) who proposed to include a vegetation model in the TF. The method is often referred to as 'model inversion' because it requires to invert the computer model to obtain reconstructed climate. Earliest versions of the method (Guiot et al., 2000; Wu et al., 2007a,b) use a statistical matching of the model outputs to pollen data ; allowing, in inverse mode, to select the most coherent climate with respect to pollen. The method assumed a static equilibrium between plant and climate, due to the use of a static vegetation model (from the BIOME model family, e.g. Cramer, 2002) and is only partly process-based due to the statistical matching. A first branch of improvements emerged with the linking of different proxies to the model outputs (Rousseau et al.,

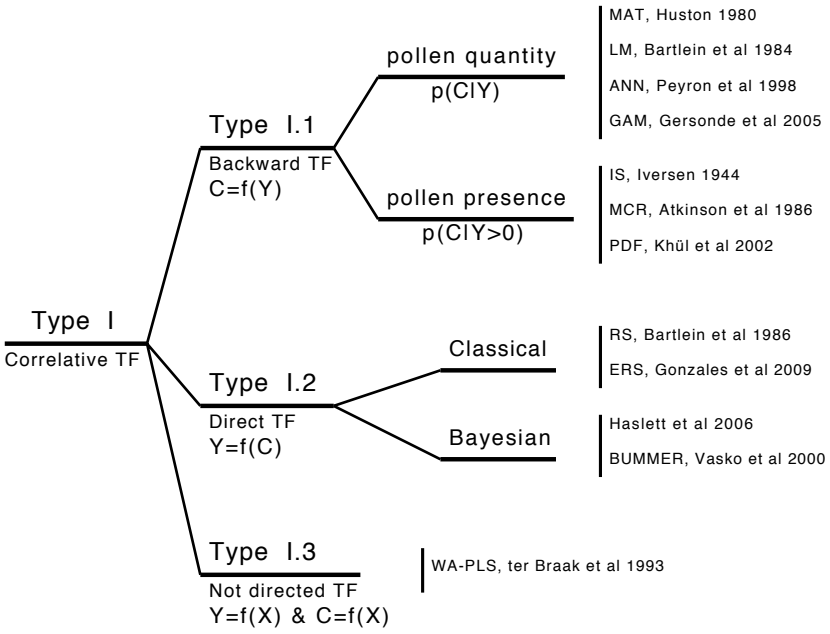

Figure 1.1: Classification of the existing correlative transfer function (type I TF). They are separated in three groups. The group I.1 contains the 'backward' TF, i.e. those modelling climate C as a function of pollen Y. They are separated in two group depending on their treatment of the pollen data. The first subgroup consider the pollen as a quantitative variable, models have the form $p(C|Y)$ (a distribution of climate given pollen data), and in the second subgroup, model only considers pollen presence, i.e. $p(C|Y > 0)$. The group I.2 contains the 'direct' TF considering vegetation as a function of climate, i.e. models of the form $p(Y|C)$. This group can be separated in 'classical' and Bayesian transfer function. The third group is formed by the WA-PLS method that does not consider a direction in its treatment of the pollen-climate link, both variables being linked to a latent (fictive) variables X.

2006; Guiot et al., 2009; Hatté et al., 2009) allowing multi-proxies reconstructions in a static and semi-process-based approach. A second branch of improvements is proposed in Garreta et al. (2009) and consists in the development of an inversion scheme for a dynamic vegetation model (LPJ-GUESS, Smith et al., 2001). It is complemented by a process-based model for the link between vegetation outputs and pollen data to achieve a full process-based approach (third and last chapter of this manuscript).

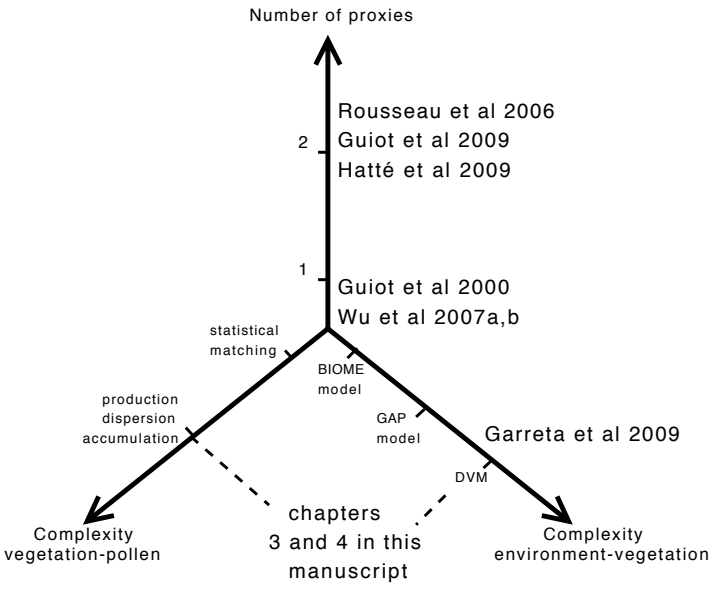

Figure 1.2: Classification of the existing process-based transfer function (type II TF). We map the various methods proposed in articles and chapters of this manuscript in a three dimensional frame. The axes of this frame are (bottom-left) the complexity or the number of processes accounted for in the vegetation-pollen relation. This complexity ranges from a pure statistical matching vegetation-pollen to a modelling of the production, dispersal and accumulation processes. (bottom-right) the complexity of the model used to model the vegetation. It ranges from the BIOME models (Cramer, 2002) to Dynamic Vegetation Models (e.g. DGVM, Cramer et al., 2001). (top) the number of proxy or variety of information used to inverse the model. It ranges from one single proxy - the pollen - to two proxies, e.g. pollen and δ^{13}C.

Since TF are 'niche' models, tools and ideas from Ecological Niche Theory (ENT, e.g. a review in Chase and Leibold, 2003) should be used to interpret, criticise and improve TF models. In palaeoclimatology, the ENT has been used as a conceptual framework helping to explain the cause of the observed species shifts, occurrences and extinctions during the Quaternary (Huntley, 1996; Jackson and Overpeck, 2000; Williams and Jackson, 2007). This framework becomes of common use for interpreting results from the

TF (e.g. Williams and Shuman, 2008) and for discussing the possible drawbacks of correlative TF (Jackson and Williams, 2004; Guiot et al., 2008). In the following section we develop the niche interpretation of the TF and show that, compared to pure niche models, all the TF used for pollen-based reconstruction have to cope with an indirect surrogate of the vegetation: the pollen sampled in sediments. This complicates the modelling by (i) creating a spatial correlation in the signal due to pollen dispersal (Telford and Birks, 2005) and (ii) providing a relative, instead of absolute, information about the vegetation composition. Both aspects are sometimes not accounted for in the TF. We propose a conceptual framework to disentangle these two aspects of the pollen-based TF by modelling on one side the climate-vegetation relation and, on the other, the vegetation-pollen relation. The process-based relation between vegetation and pollen has been subject of discussion since von Post (1916), Tauber (1965) and Webb (1974). Today, several models exist (e.g. Sugita, 2007a; Paciorek and McLachlan, 2009) that could be scaled to account for continental scale pollen dispersion (objective of the third chapter in this manuscript).

In ecology, a pressing need for understanding and predicting future distribution of species under climate change requires the calibration, validation and use of Species Distribution Models (SDM). These models, very close in spirit to the TF raise the same challenges as in palaeoclimatology. Since conceptual and modelling frameworks are lacking a consensus, the dual vision proposed by correlative versus process-based models increases attention on the underlying hypothesis behind the correlative models and raises interest for the process-based approach (Kearney, 2006; Morin and Lechowicz, 2008; Wiens et al., 2009). We propose a framework that shows the direct analogy between reconstructing past climate from pollen and predicting future plant species distributions. These fields of research are derived from palaeoclimatology and ecology, two disciplines that are sufficiently different to provide complementary visions and questionings about the same objects: niche models, i.e. niche theory representations and applications. We believe that it will increase the interest for novel modelling approaches by extending their potential range of application.

The next section starts with a brief mathematical classification and review of the correlative TF. In section 1.2.2, we present a framework to disentangle the two major aspects of the climate-pollen relation which are the niche model for the climate-vegetation model and the processes linking vegetation and pollen. In section 1.2.3 we discuss the hypotheses and constraints inherent to the correlative approach for niche modelling compared to process-based modelling. In section 1.3 we present a Bayesian hierarchical framework to unify the niche modelling approaches in pollen-based palaeoclimatology and future plant distributions in ecology.

1.2 TF in palaeoclimatology

1.2.1 Review of the correlative TF

Correlative TF are models of a link between, $Y_i = (Y_i^1, .., Y_i^k)$ a pollen assemblage formed with k taxa at a site i, and a set of l climate variables $C_i = (C_i^1, .., C_i^l)$ at the same site. The reconstruction process is twofold. The *calibration* step consists in inferring the parameters of the TF based on a modern dataset of climate and pollen $(c, y)_{s=1..N}$. This dataset must be spatially distributed (at a continental scale) and massive ($N > 1000$, e.g. Whitmore et al., 2005; Williams et al., 2006). This is imposed by the need to infer a robust link between climate and pollen over a large climate range. The *reconstruction* step consists in obtaining information about the climate C_t based on fossil pollen y_t.

In the type I.1 TF based on the Presence/Absence (PA) of taxa (IS, MCR and PDF), the distribution of a small set of climate variables is modelled given that the species is present. They have statistical models of the form

$$p_\theta(C_i|Y_i) = \prod_{j=1}^{k} p_{\theta^j}(C_i|Y_i^j > 0) \tag{1.1}$$

with $p_{\theta^j}(C_i|Y_i^j > 0)$ a distribution of the climate variables given that the jth taxa at site i is present. In the IS and MCR, these distributions are uniform distributions and in the PDF they are bivariate Gaussian distributions. The calibration consists of obtaining the contours or parameters for the climate distributions based on a modern dataset that may include any data about species PA, generally, vegetation atlases. Due to their backward form (climate is a function of the PA), the reconstruction is defined in statistics as a simple *prediction*. The predicted climate distribution is the intersection (normalised product) of the climate distributions associated with the taxa present in the fossil pollen.

The attraction of the backward approach is that it directly provides reconstructions as simple predictions. Several statistical models have been used to exploit this feature, modelling the climate response to pollen assemblage using more and more complex regression models, e.g. linear model, GAM, ANN. These TF have the form $C_i = f(Y_i, \theta) + \epsilon_i$, where $f(.)$ is a function depending on the pollen Y and parameters in θ and an errors term ϵ_i. The calibration consists in fitting the parameters to a value $\hat{\theta}$ with caution against over-fitting (e.g. techniques in Ripley, 1996). The climate reconstruction C_t for a pollen assemblage y_t is readily obtained by plugging y_t into the fitted function, i.e. as $C_t = f(y_t, \hat{\theta})$. Two levels of uncertainties could be readily included in the confidence interval of this prediction: (i) the errors ϵ and (ii) the uncertainty on $\hat{\theta}$ fitted in calibration. But it is unclear how the confidence interval for the reconstruction are provided (e.g. early uses of these methods, Bartlein et al., 1984; Peyron et al., 1998; Gersonde et al., 2005).

In the MAT, a kernel shape and a bandwidth (corresponding in the MAT to (a) the distance metric and weighting of the point, and (b) the maximum distance for selection) are selected using cross-validation. These kernels and bandwidth allow the smoothing of climate in a pollen space. Then, reconstruction for a pollen assemblage y_t is provided as the local smoothing of climate around y_t (Figure 1.3). The classical methods for providing confidence intervals in the context of local smoothing are replaced in the MAT by the variance of the selected analogues (Guiot, 1990) or, the errors measured in a leave-one-out cross validation performed on the modern dataset

(Nakagawa et al., 2002). These methodologies are criticised (Telford and Birks, 2005; Telford, 2006; Telford and Birks, 2009) because a strong autocorrelation in the modern pollen and climate variables hides the potential variability in the analogues. Tools from local regression literature (e.g. Loader, 1999) may improve confidence intervals by allowing inclusion of the autocorrelation present in the pollen and climate modern data.

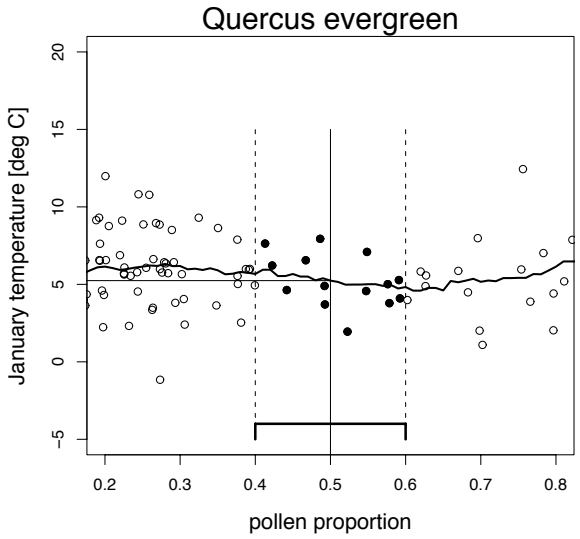

Figure 1.3: Interpretation of the Modern Analogue Technique (MAT) as a local smoothing of climate in a space defined by the pollen composition. Here we use a single taxa - Quercus evergreen - and climate variable - January temperature - to illustrate the method. The points on the graph are the proportion of the pollen type versus the January temperature at the 1301 sites of the European dataset used in this manuscript (see next chapter for the dataset description). For a fossil pollen abundance of 0.5 (vertical central line) all the points whose Euclidian distance is lower than 0.1 (black dots between the dashed lines) are retained as Analogues. The reconstruction is provided as the means of selected points (horizontal line). This corresponds to a smoothing with the kernel shown on the bottom of the graph, whose smoothing over the whole range of pollen proportion is the dark, smoothly evolving curve.

In type I.2 TF, pollen is expressed as a function of climate. These TF can be seen as

independent mappings of the pollen proportions on a small set of climate variables, i.e. projections of taxa proportions on the space defined by climate variables. Pollen data is most often available under the format of relative frequencies, called 'multinomial' data in the case of raw counts (Mosimann, 1963) and 'compositional' data (Aitchison, 1982) in the case of proportions or percentages. Their proper modelling thus requires to consider all the proportions of an assemblage at the same time through a sum-to-one constraint (the sum of the taxa proportions is constrained to one). In the RS and ERS methods, the response models (surfaces) are independently fitted for each pollen proportion. The sum-to-one constraint is added *a posteriori*. This precludes the proper quantification of the uncertainties and potentially induces a misinterpretation of the models' parameters (Aitchison, 1982). These TF are based on the model

$$p_\theta(Y_i|C_i) = \prod_{j=1}^{k} p_{\theta^j}(Y_i^j|C_i) \tag{1.2}$$

with $p_{\theta^j}(Y_i^j|C_i)$ a model of the pollen taxa j at site i given climate C_i. These models are parametrical or not and their calibration is performed by fitting independently a pollen response surface per taxa. In a Maximum Likelihood framework (ML, one of the classical inference frameworks, e.g. Young and Smith, 2005), the reconstruction from 'direct' TF is not defined since climate C is considered as a fixed and known regressor. The inversion of the model for reconstruction is obtained using a heuristic algorithm based on the MAT method described before. Reconstructed climate C_t for the pollen sample y_t is the one associated to the point of the response surface that best matches y_t (Bartlein et al., 1986). This technique approximates C_t but does not define a confidence interval. The Bayesian framework (e.g. Robert, 2001; Gelman et al., 2004) provides the theoretical and technical tools for defining and effectively obtaining in a consistent way the reconstruction with its associated confidence interval.

The other type I.2 methods (BUMMER and the method in Haslett et al., 2006) are based on proper models for the pollen data as relative frequencies and they adopt a Bayesian framework for inference, solving the problems of the reconstruction definition, uncertainty propagation, and confidence interval definition. The core structures of these model are analogous to those of the previous response surfaces. BUMMER uses

Gaussian curves for the species responses in the spirit of symmetric unimodal distribution (ERS) and the Haslett et al. (2006) model uses semi-parametric surfaces (RS). We group them under the general form

$$p_\theta(Y|C) = p(Y_i|p_i) \prod_{j=1}^{k-1} f_{\theta^j}(t^j(p_i)|C_i) \tag{1.3}$$

with $p(Y_i|p_i)$ a distribution for the data (Dirichlet-multinomial in both cases), $t(.)$ a transformation that maps p_i in an unconstrained space of dimension $k-1$ and f_{θ^j} is either (i) a Gaussian curve relating the jth transformed component $(t^j(p_i))$ with climate C_i in BUMMER or (ii) a smooth random field of $t^j(p_i)$ in a climate space (Haslett et al., 2006). Bayesian inference for calibration and reconstruction of past climate from these models is not straightforward and requires Markov Chain Monte Carlo algorithms (MCMC, Robert and Casella, 1999). In the case of the Haslett's model, such an algorithm is so computationally demanding that it delays the wide use of the method. Despite this problem requiring further development in statistics, we believe that these Bayesian approaches form the next generation of correlative TF because their modelling is proper and allows to define 'confidence intervals' under the form of 'posterior distributions', i.e. reconstructions are provided in the form of distribution whose median, mean, standard deviation etc. can be computed.

1.2.2 TF are niche coupled with vegetation-pollen models

The 'direct' TF (type I.2 and II) describe the pollen response to a few climate variables. Therefore, rather than just modelling niches - the climate-species relation - they also have to cope with the pollen, a proxy for the species. This makes them models of plants climatic niches embedded in models of the pollen accumulated in sediments as a function of vegetation. We build a two-piece model based on previously published models for the vegetation-pollen and environment-vegetation relationships. This illustrates the need to explicitly consider the pollen-vegetation relation as part of the TF, mainly by modelling a spatial correlation in the pollen data.

Let introduce a simplified relation between pollen and vegetation as those used,

for example in Prentice and Parsons (1983); Sugita (2007a); Paciorek and McLachlan (2009) and the third chapter of this manuscript,

$$Y_i = f^1(V_i, V_{\bar{i}}) + \epsilon_i^1$$

with Y_i the pollen data sampled at site i, f^1 a deterministic or stochastic function with inputs V_i, the vegetation at site i and $V_{\bar{i}}$ the vegetation surrounding (several km around) site i. ϵ_i^1 is a site-specific error supposed additive representing the unpredictable part of the pollen signal. The processes that should be modelled in f^1 (Davis, 1963; Webb, 1974; Prentice and Parsons, 1983; Prentice, 1985; Sugita, 2007a) include pollen production, dispersal, and accumulation in the lake sediments. The pollen dispersal, sometimes said 'long distance' in palynology, allows - and imposes - to link pollen at site i with the vegetation surrounding the site, i.e. including $V_{\bar{i}}$ in the relation.

On another side, a simple species-climate niche model would relate vegetation species (or taxa) to their niche variables (e.g. climate) following the relation

$$\mathrm{V}_i = f^2(C_i) + \epsilon_i^2$$

with f^2 a deterministic or stochastic function with input C_i, the niche variables considered. ϵ_i^2 is a site-specific error supposed additive which represents the unpredictable part of the vegetation. This model may include spatial correlation to translate a vegetation patchiness but we keep the problem simple and conservative in the sense of not overstating spatial correlation.

The coupling of both models, thus creates the following structure

$$Y_i = f^1\left(f^2(C_i) + \epsilon_i^2, \sum_{s\in\bar{i}} f^2(C_s) + \epsilon_s^2\right) + \epsilon_i^1 \tag{1.4}$$

where $\sum_{s\in\bar{i}} f^2(C_s)+\epsilon_s^2$ is the sum of the species-climate relation applied to all the points in the domain $\bar{i}$. For a given climate C over all the points in space ($C = C_{i=1..N}$ e.g. modern climate), the pollen collected at two different, but 'close' sites, Y_i and Y_j co-vary due to the common error term $\sum_{s\in\bar{i}\bigcap\bar{j}} \epsilon_s^2$. The magnitude of this spatial covariance is given by the amplitude of noise (ϵ_1 and ϵ_2) and the closeness of modern sites relative to pollen dispersal distance. The amplitude of noise depends on the model structure

but pollen dispersal distance, which is different for each taxa and ranges between ten to hundred kilometres (see e.g. the special issue Gaillard et al., 2008, and references therein) is of the same order as distances between modern pollen sites (see Figure 1.4).

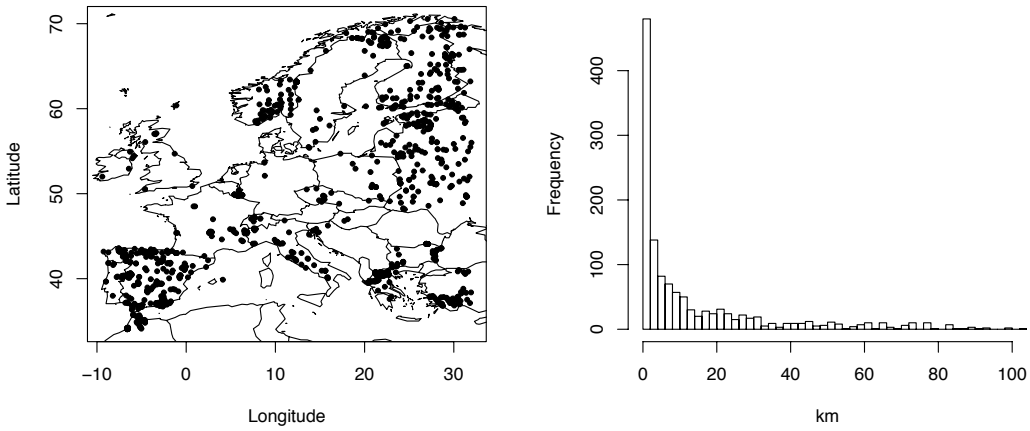

Figure 1.4: Spatial repartition and distances distribution of the 1301 modern pollen sites used in this manuscript (see next chapter for a description). (left) Spatial repartition of the pollen surfaces samples over Europe. (right) histogram of the distances between each point and its nearest neighbour. More than 75% of the points have their nearest neighbour at less than 20km.

This spatial correlation is never modelled nor tested in classical TF despite it creates an overconfidence in the results if not taken into account (Legendre, 1993; Telford and Birks, 2005). Thompson et al. (2008); Telford and Birks (2009) propose to use *h-block* sampling (Burman et al., 1994) to properly estimate the calibration uncertainties. This does not allow to correct estimation bias which would be avoided only by modelling the spatial correlation (allowing proper error quantification at the same time). Such modelling could be achieved in a process-based fashion by re-using the relations in Prentice and Parsons (1983); Sugita (2007a); Paciorek and McLachlan (2009) or in a geostatistical fashion by specifying a spatial correlations structure for the residuals. The numerical complexities arising from considering a spatial structure, i.e. a joint structure

for all points at the same time, is very constraining for inference (e.g. third chapter in this manuscript or Paciorek and McLachlan, 2009) since pollen data are numerous and multivariate (several taxa sampled at thousands of points). And this is likely the same for a geostatistical treatment (Higdon, 1998; Fuentes, 2007; Stein, 2008; Cressie et al., 2007; Zhang and Wang, 2009). These problems make the modelling and inference of an improved type I.2 TF as challenging as a full process-based approach involving a computer model inversion.

The process-based approach is historically based on the use of a computer model for the vegetation-climate link (Guiot et al., 2000). This makes natural the dichotomy between climate-vegetation and vegetation-pollen relations in the TF. In the first approaches (Rousseau et al., 2006; Wu et al., 2007a,b; Garreta et al., 2009; Hatté et al., 2009), authors used a statistical approach for the vegetation-pollen relation that is based on the idea of 'matching' or correspondence between model outputs and pollen samples. These approaches do not use spatial relations between the vegetation and the pollen. In the third and last chapters of this manuscript we propose a model and try the inference of pollen dispersal for the vegetation-pollen link; forming a full process-based TF. As for the type I.2 TF, this complicates the inference and requires statistical approaches still in their infancy.

1.2.3 Intrinsic aspects of the correlative and process-based approaches

In this section we review the intrinsic hypotheses behind the correlative TF seen as niche models (Chase and Leibold, 2003) and exclude the problem of the vegetation-pollen link modelling already discussed. In this perspective, the type I.2 (direct) TF are directly interpretable. However, the remarks also apply to the type I.1 and I.3 (backward and not directed) TF that are less straightforwardly or not interpretable in term of the processes generating their niche.

The definition of a species niche is at the heart of niche theory and thus in constant evolution. We use a modern definition (Chase and Leibold, 2003) of the niche as 'the environmental conditions that allow a species to satisfy its minimum requirements so that the birth rate of a local population is equal to or greater than its death rate along with the set of per capita impacts of that species on these environmental conditions'. This definition recognises, at least, two main characteristic points to the niche. (i) It is dual, i.e. is defined in terms of 'requirements' and 'impacts', (ii) it is dynamical ('the birth rate (...) is equal to or greater than') allowing the populations to have their own dynamics.

Low and arbitrary dimension of the niche

The correlative (at least type I.1) TF, that are descriptive models of the plant species niches are defined on small and *a priori* (in the sense of derived from theory instead of statistically selected) set of climatic-only variables raising questions about the reliability of their description and therefore of their skills for palaeo-climate and environment reconstruction.

As stated by the precedent definition, the niche dimension, i.e. the number and type of 'environmental conditions' are a major, if not the single, parameter defining the niche. The environmental variables having a potential role in the niche of a plant species include,

- climate variables (e.g. temperature, precipitation, sunshine),
- other physical environmental variables (e.g. CO_2, insolation),
- biotic factors (e.g. soil type, nutrients and water availability),
- predators, competitors and facilitators (e.g. fauna components, fungi, other plants competing for light)

In the building of a descriptive TF, the niche variable selection (niche dimension definition) should be the main task of interest. To be as objective as possible, the variable

selection should be statistical, for example using the variable selection techniques used in multivariate regression analysis. Moreover, the selection can only be performed in the set of conditions expressed in the modern pollen-climate datasets.

We focus on two examples to illustrate that, by considering a potentially too restricted dimension of the niche, the TF may lead to biased or overstated palaeoclimate reconstructions.

Atmospheric CO_2 is known to have a major effect on the plant response to climate (e.g. review in Prentice and Harrison, 2009), principally by controlling (1) the efficiency of C_3 plant photosynthesis and (2) their drought stress resistance. Being a well-mixed gas in the atmosphere, the CO_2 modern concentration does not vary significantly in space to be accounted for in the correlative TF. Since this modern concentration (from around 310ppmv in 1950 to more than 385ppmv today) does not overlap the range of its values during the recent glacial-interglacial periods (170 to 300ppmv), plant-climate relations fitted on modern dataset are potentially distorted for past climate reconstructions. The problem has been discussed around the paper of Cowling and Sykes (1999). Quantification of the errors made by neglecting CO_2 in TF are based on the use of vegetation models (Jolly and Haxeltine, 1997; Wu et al., 2007a,b). Wu and coauthors found differences of around 10°C for certain regions when taking or not CO_2 variations into account at the Last Glacial Maximum (LGM). The only solution to this problem is the use of a process-based approach exploiting relations between plant and CO_2 that have been calibrated on laboratory experiments and that are available, for example in vegetation models.

Competition between plants for limited resources is a major process expected to generate forest structure and forms the basis for many vegetation models (e.g. GAP models, Bugmann, 2001). In the direct TF it is clear that inter-species interactions are not modelled in 'response surfaces' that are fitted independently per taxa nor in their residuals. This creates an overconfidence in the results recently detected in cross-validation for the Haslett et al. (2006) TF that provides too narrow confidence intervals (Salter Townshend, 2009, and John Haslett, personal communication). A process-

based approach based on the use of a GAP model includes, for example, competition for light. In GAP or other models, such description of competition processes will always be incomplete and that's why the use of a mechanistic vegetation model inside a TF must always come with a modelling of the mechanistic-model structural errors.

Steadiness of the correlative TF

The absence of dynamic vegetation modelling in classical TF challenges their reliability in two ways. First, depending on the rapidity of the plant response to climate change, climate can be tracked more, less, or not accurately using pollen records (Webb III, 1986; Prentice, 1986, 1988). Second, the slower the plant spatio-temporal response is, compared to climate change, the more vegetation composition at a time t is dependent on its history; precluding the direct interpretation of a single time point in a chronology because it could result from different climates due to different historical vegetation. Since the works of Prentice et al. (1991) it has been accepted - sometimes as an unavoidable hypothesis - that, at the centennial to millennial scale of pollen-based reconstructions, vegetation is in 'dynamic equilibrium' with climate.

The hypothesis of 'dynamic equilibrium' ('vegetation response is fast enough to have kept up with the changes', Prentice et al., 1991) may be understood and agreed at the centennial to millennial scales. But today, climate is rapidly evolving in its mean and extreme events, and CO_2 increase is unprecedented in the last glacial cycles, at least since 1950 (Solomon et al., 2007). This makes probable that modern vegetation composition used to calibrate the TF is not in the equilibrium required to properly fit the pollen-climate relation.

In the reconstruction of climate from low-resolution sediment cores (when samples are separated by large time steps) the equilibrium hypotheses is acceptable since the timing of climate change events is, anyway, poorly constrained. When high-resolution cores are used the temptation is high to interpret vegetation changes - reconstructed independently between core points - as a synchronous climate change (but see studies assessing vegetation-climate lag properly: Ammann et al., 2000; Williams et al., 2002).

The proper timing of past climate change is clearly not possible from pollen records at a single site when using correlative TF that are static. In the process-based approach, we propose to take the response delay into account by modelling a dynamic response of the vegetation to climate change using a Dynamical Vegetation Model (DVM, in this manuscript LPJ-GUESS, Smith et al., 2001).

Descriptive and process-based approaches for extrapolation

The problem of unreliability of the correlative TF outside the range of the modern conditions is known in palaeoclimatology under the name of 'no-analogue' problem. In statistics, it is a classical extrapolation problem caused by the TF that are not process-based and worsened in the case of the not parametrical TF. Any attempt to reduce this problem with correlative TF require to assume new hypotheses that are not testable. For example, Vasko et al. (2000); Gonzales et al. (2009) proposed to extend a response surface model outside its range of modern calibration by assuming either a Gaussian or symmetric unimodal response to climate by taxa. This choice of a model for the whole 'niche' shape emerges from high-level considerations on plant ecology that are not directly testable on real dataset despite they have a major impact on palaeoclimate reconstruction. On the opposite, process-based TF lie on the representation of basic and sometimes directly measurable processes. When such processes are based on physical laws, their extrapolation outside the rage of calibration (if sufficiently large to allow a 'good' calibration) has a strong scientific justification.

1.3 A framework to share knowledge between ecology and palaeoclimatology

In ecology, Species Distribution Models (SDM, e.g. Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005; Elith and Leathwick, 2009; Kearney and Porter, 2009) are the equivalent of the 'direct' transfer functions used in palaeoclimatology. These models are used to calibrate a species response to its environment from the modern, spatial dis-

tribution of species and climates. Calibrated on large, spatial, datasets, they are used to infer the future species distributions under changing climate. As in palaeoclimatology, interest for process-based SDM is increasing and their comparison with correlative approaches allow critical insight into the underlying hypotheses of each (Morin and Lechowicz, 2008; Kearney and Porter, 2009). SDM are strongly based on a niche theory interpretation (Wiens et al., 2009) and they use the most up-to-date tools in terms of geographic information systems (GIS), statistical inference methods and modelling (Elith and Leathwick, 2009). The process-based approaches are under development and are either based on the use of vegetation models (e.g. Morin and Thuillier, 2009) or on statistical relations copying the processes expected to relate species to their environment (Kearney and Porter, 2009).

The general methodology for SDM is the same as for TF (see Figure 1.5 for a schematic illustration). The model is calibrated using a modern, typically as massive as possible and spatial dataset. Many tests should be realised to select the variables defining the niche, assess the model accuracy etc. In a second step, the model is used to predict the species distributions under different climate projections. The analogy between models in palaeoclimatology and ecology becomes evident following the remark we made section 1.2.2; TF should be SDM coupled with vegetation-pollen models. Even if these models have the same structures, they must match in their scales and applications to be readily transferable. Indeed, the dominant processes can be different for different scales and/or their controlling parameters can vary. Time scales of Quaternary climatology, linked to pollen data resolution, are roughly on the order of tens of years for the highest quality records. The projections of the future species distributions are made on shorter time scales, for the next decades but the magnitude of changes expected for the future is potentially the same or higher than in the last glacial cycles. Spatial scales span the range of local (one forest) to continental for both palaeoclimatology and ecology.

Both approaches of the species distributions (SDM and TF) are lacking a consistent framework for handling the calibration uncertainties, i.e. uncertainties are not propa-

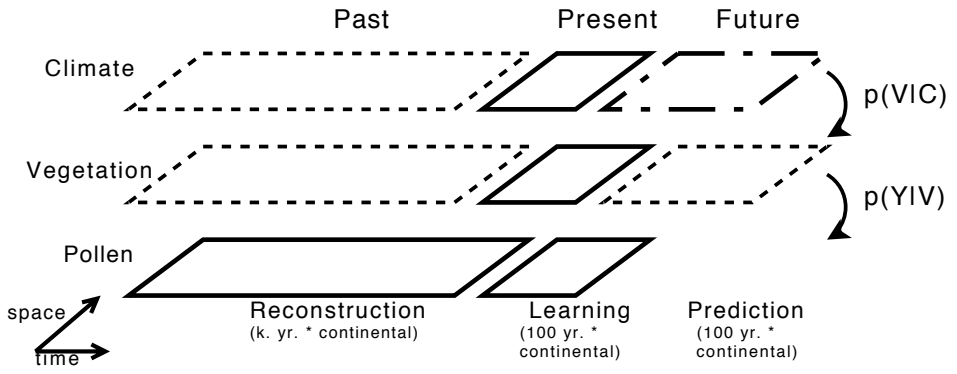

Figure 1.5: Representation of the spatio-temporal dimensions involved in the classical SDM and TF modelling and use. (left) The 'past' period for pollen-based palaeoclimate reconstruction is registered in pollen samples covering several tens of thousand years (k yr.) and registered in lakes distributed at a continental scale. The palaeoclimatology challenge is to obtains vegetation and climate reconstructions at these scales. (center) The modern period is defined by the length of time for which we have instrumental records of climate, vegetation and pollen. It is several decades to centuries long. In this period we measure and continue to monitor pollen, vegetation and climate at the scale of the earth. It allows to learn about the relations linking climate, vegetation and pollen that are used to reconstruct or predict the variables one from the others. (future) It is the period for which we need prediction. The political and economical interest is focused on the next tens or fifty years. In this period we have more or less credible scenarios for climate at a continental scale. The challenge is to predicted the range of possible vegetation change from these scenarios.

gated from the calibration to the prediction/reconstruction. These uncertainties can be on the model's parameters or on the models themselves if different niche models are competing. For palaeoclimatology we discussed this point previously. In ecology the same questions are raised (Elith and Leathwick, 2009). We propose to dive the TF (direct and process-based) and SDM (correlative and process-based) in a Bayesian framework. This,

- reinforces the analogy between TF and SDM,

- provides a consistent framework to characterise and propagate uncertainties between calibration and reconstruction/prediction in both disciplines,
- may allow to exchange tools between disciplines.

The Bayesian paradigm has several advantages for palaeoclimatology as discussed in the case of the direct TF. In this framework, all the quantities of interest (e.g. climate, vegetation, pollen) are seen as random, i.e. no fundamental distinction is made between parameters, dependent variables, independent ones, etc. The modelling consists in specifying distributions linking the random quantities and the inference consists in obtaining the distribution of the quantities of interest given the model structure and the data available. Let describe a general calibration and reconstruction/prediction framework.

Following the remark of section 1.2.2 (slicing between climate-vegetation and vegetation-pollen models), a proper pollen climate link could be modelled as

$$p_0(Y|V,\theta_0)\, p_1(V|C,\theta_1)$$

with C the values of the niche variables selected. Potentially climate but other variables can be included. $p_0(Y|V,\theta_0)$ is a pollen-vegetation link with parameters θ_0 and $p_1(V|C,\theta_1)$ a vegetation-environment link with parameters θ_1. p_1 could be either a statistical model (e.g. type I.2 TF), a deterministic vegetation model (Guiot et al., 2000) or a stochastic one (Garreta et al., 2009).

The calibration consists in inferring the parameters θ_1 and θ_2. Suppose we have a classical pollen and environmental dataset at a continental scale, noted $(y,c)_s$ with s demoting a large set of N sampled sites. Following the Bayes formula, calibration consists in obtaining the *posterior* distribution

$$p(\theta_0,\theta_1|y_s,c_s) \propto \int p_0(y_s|V,\theta_0)\, p_1(V|c_s,\theta_1)\, p(\theta_0,\theta_1)\, dV$$

with $p(\theta_0,\theta_1)$ a *prior* distribution translating the amount (or absence) of knowledge we have on the parameters θ_0 and θ_1. Effectively obtaining this distribution is a hard

problem due to its high dimension (equal to $\dim(\theta_0) + \dim(\theta_1)$) and the integral over the vegetation field (V is vegetation at all sites). Haslett et al. (2006) and the third chapter of this manuscript use MCMC algorithms allowing to obtain nearly every type of distribution under the form of simulations of θ_0 and θ_1.

The climate reconstruction for a single pollen sample y_t is obtained through the Bayes formula as

$$p(C_t|y_t) \propto \int p_0(y_t|V_t, \theta_0)\, p_1(V_t|C_t, \theta_1)\, p(\theta_0, \theta_1|y_s, c_s)\, p(C_t)\; dV_t\, d\theta_0\, d\theta_1$$

with $p(C_t)$ translating the prior information we have on the environment at time t before reconstruction. This formula includes - explicitly - the uncertainties on the parameters obtained through calibration which is not presently done in most of the TF. The distribution can be effectively obtained using, e.g. a MCMC algorithm.

The Bayesian calibration of SDM could involve exactly the same kind of model as in the TF, i.e.

$$p_2(Y|X, \theta_2)\, p_3(X|C, \theta_3)$$

with $p_2(Y|X, \theta_2)$ a link between the data Y that can be noisy surrogates of the species X, with parameters θ_2 and $p_3(X|C, \theta_3)$ a species-environment link with parameters θ_3. p_3 could be either a statistical model (e.g. correlative SDM, type I.2 TF), a deterministic vegetation model (Morin and Thuillier, 2009) or a stochastic one (Garreta et al., 2009). Calibration thus consists in obtaining $p(\theta_2, \theta_3|y_s, c_s)$ as before, based on a set of modern data composed of the species indicator and its environment recorded on a set of $s = 1..N$ sites.

The prediction of the species under possible climate for time t (in the future) is obtained following the Bayes formula

$$p(X_t) = \int p_3(X_t|C_t, \theta_3)\, p(\theta_2, \theta_3|y_s, c_s)\, p(C_t)\; d\theta_3$$

with $p(C_t)$ the distribution of possible future climate that is often provided as a set of simulations from different global circulation models (referred to as a climate ensemble simulation). This formula includes explicitly the uncertainties on the parameters

obtained through calibration which is not done in ecology (e.g. Elith and Leathwick, 2009). Note that Bayesian theory provides tools to readily expand such formula to also include a probability of model if several models $p_3(X_t|C_t, \theta_3)$ are competing for the link species-environment.

1.4 Conclusion

More than sixty years after the first backward approach of Iversen (1944) and thirty years after the first direct approach of Imbrie and Kipp (1971), the statistical and computational tools are so elaborated that they start to allow the realisation of fully process-based palaeoclimate reconstruction, which includes decades of works in many disciplines around biology and ecology. This unique opportunity to provide a new and more accurate picture of past climate and environment has to be developed. The slow startup of process-based palaeoclimate reconstruction after the work of Guiot et al. (2000) may be explained by the high level of expertise in statistics, stochastic inference and computation it requires. We believe that more interactions between statisticians and palaeoclimatologists will help in its development.

In bringing closer palaeoclimatology and ecology we believe that both field will enrich and challenge the other one. From its side, ecology stands on a constantly evolving theoretical representation of the species-environment interactions which helps in reconsidering the models used for their description. On its side, palaeoclimatology has unique datasets to test models for large scale spatio-temporal dynamics. Even though the palaeoclimatological datasets are incomplete (they lack most environmental conditions) they contain the remains of many past dynamics of life and must be used when we will need to calibrate realistic models of the future biota dynamics. Rephrasing Jackson and Williams (2004), we believe that in developing fully process-based approaches in close cooperation with ecologists, the past no-analogues will become keys for the future ones.

Chapter 2

A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model

This chapter is published online in *Climate Dynamics* as the following research paper

Garreta, V., Miller, P. A., Guiot, J., Hély, C., Brewer, S., Sykes, M. T., and Litt, T. (2009). A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model. *Climate Dynamics*, doi: 10.1007/s00382-009-0629-1

I realised the work presented in this chapter with the help of Paul Miller (Geobiosphere Science Center, Lund University, Sweden) for the development of LPJ-GUESS code. I wrote the chapter except the LPJ-GUESS description section by Paul.

Abstract Climate reconstructions from data sensitive to past climates provide estimates of what these climates were like. Comparing these reconstructions with simulations from climate models allows to validate the models used for future climate prediction. It has been shown that for fossil pollen data, gaining estimates by inverting a vegetation model allows inclusion of past changes in carbon dioxide values. As a new generation of dynamic vegetation model is available we have developed an inversion method for one model, LPJ-GUESS. When this novel method is used with high-resolution sediment it allows us to bypass the classic assumptions of (1) climate and pollen independence between samples and (2) equilibrium between the vegetation, represented as pollen, and climate.

Our dynamic inversion method is based on a statistical model to describe the links among climate, simulated vegetation and pollen samples. The inversion is realised thanks to a particle filter algorithm. We perform a validation on 30 modern European sites and then apply the method to the sediment core of Meerfelder Maar (Germany), which covers the Holocene at a temporal resolution of approximately one sample per 30 years. We demonstrate that reconstructed temperatures are constrained. The reconstructed precipitation is less well constrained, due to the dimension considered (one precipitation by season), and the low sensitivity of LPJ-GUESS to precipitation changes.

2.1 Introduction

Numerous studies have produced statistical palaeoclimate estimates by using the modern relationship between pollen and climatic data (e.g. the pioneering works of Webb and Bryson (1972) and Prentice and Helmisaari (1991) or a recent review in Guiot and De Vernal (2007)). These studies have substantially improved our knowledge of past climates and have been used as benchmark to evaluate robustness of climate models (e.g. from COHMAP Members (1988) to Jost et al. (2005)).

The existing reconstruction methods are based on the assumption that plant-climate interactions remain the same through time, and implicitly assume that these interactions are independent of forcings such as changes in atmospheric CO_2. Guiot et al. (2000) showed that this assumption could produce significant biases in the results and that by using a vegetation model inversion, it was possible to evaluate these biases and to correct them. Wu et al. (2007a) applied the method to European, African and Asian data for two periods of the past when atmospheric CO_2 concentration was significantly different from the present one. They showed that biases could reach up to 10°C for winter temperature in Europe during the Last Glacial Maximum. These papers used an equilibrium vegetation model (BIOME4) which accounts for processes related to carbon and water cycles, but not for those related to plant competition and mortality. A more recent and sophisticated dynamic model, LPJ-GUESS (Smith et al., 2001), takes these processes into account.

We propose a method for the inversion of a dynamic vegetation model and argue that in palaeoclimatology this method is an improvement compared to the inversion of static models. Indeed, changing the static vegetation models used previously for a up-to-date dynamic model updates the transfer function defined by inversion. Second, when the dynamic inversion (read inversion of a dynamic model) is applied to a high time-resolution sediment core it provides a way to bypass the classic assumptions of:

- *independence between samples.* With the exception of the Haslett et al. (2006) method, classic reconstruction methods ignore temporal correlations even if pollen

data are sampled in sediment cores, which provide temporal records of the pollen. The dynamic vegetation model simulates vegetation histories that can be used as a natural link for temporal reconstruction.

- *equilibrium between climate and vegetation.* Classic transfer functions are calibrated with modern pollen and climate data which are necessarily spatial. The absence of any temporal information means that we cannot calibrate a dynamic link or disequilibrium. Under a changing climate (modern or past) this is a simplification because the vegetation response may be delayed. Using LPJ-GUESS to simulate vegetation dynamics allows us to include a delay between climate change and vegetation change, by taking into account growth, mortality and competition processes.

Both of these assumptions are admissible when working with a low time resolution because the expected reconstruction is of low resolution and samples are nearly independent when there is a long time interval between samples. When a high resolution core is used, the expectation is a high quality reconstruction, i.e. including and properly quantifying all possible sources of uncertainty. In this case, the noise associated with the independent reconstruction of samples should be reduced or properly quantified by modelling a link between samples. The equilibrium hypothesis must also be considered, as this potentially induces error in the timing of climate changes.

The inversion of the dynamic model LPJ-GUESS, compared to static model inversion, is complicated in two ways.

First, LPJ-GUESS is stochastic, which means that any two simulations realised with the same forcings (climate, CO_2, soil, etc) are not exactly identical. This is due to fire, establishment and mortality processes which are represented in LPJ-GUESS as stochastic. The vegetation must therefore be considered as a random "hidden" variable instead of a deterministic function of climate as in Guiot et al. (2000).

Second, we want to use the temporal aspect provided by the vegetation model. In the reconstruction algorithm, this would require a high-dimensional climate space to

be tested and induces the classic problem of the "curse of dimensionality". For a static model, we can run the model for a single point in time. Here several possible climates are proposed, the model is run with each scenario and the simulated vegetations are compared to a single pollen sample to retain the more coherent simulations (Guiot et al., 2000). With the same method for temporal inversion, we need to propose several high-dimensional climate histories, simulate several vegetation histories and compare them to the pollen history. This algorithm is inefficient and requires massive simulations because almost no vegetation chronology will fit the entire pollen chronology. With this kind of algorithm (a global stochastic search in the entire time-climate space) computing time for simulation is prohibitive, at least due to the use of a vegetation model.

To overcome both challenges and perform the temporal inversion, we have developed, and present here, a hierarchical Bayesian model and a particle filter algorithm (Doucet et al., 2001) for inference. The hierarchical Bayesian approach facilitates the probabilistic formalisation of the inversion process. It has general attractive features in paleoclimatology (see the discussion in Haslett et al., 2006) and we mainly use its concept of organisation of the variables into a hierarchy, *prior* and *posterior* described in the modelling section. The particle filter algorithm is mainly a Bayesian tool used for inference in real-time (or on-line) problems. In our context it allows to bypass the curse of dimensionality because it considers the reconstruction date after date.

The paper is structured as follows: (i) The vegetation model, climate data and pollen data are presented. (ii) We then describe the statistical model and inference algorithm. (iii) The method is validated using 30 modern pollen samples distributed across Europe. (iv) The temporal feature of the approach is fully exploited by reconstructing Holocene climate from the high resolution Meerfelder Maar sediment core (data from Litt et al., 2009).

2.2 Materials and methods

2.2.1 The LPJ-GUESS Dynamic Global Vegetation Model

LPJ-GUESS simulates the dynamics of vegetation stands, accounting for competition between tree individuals and populations as a forest gap model (Shugart, 1984; Bugmann, 2001). A full description of the model can be found in Smith et al. (2001). Biophysical and physiological processes are represented mechanistically, and are based on the same formulations as the well-evaluated Lund-Potsdam-Jena dynamic global vegetation model (LPJ-DGVM; Sitch et al., 2003). Updates to the model's hydrological processes were described by Gerten et al. (2004).

In LPJ-GUESS, cohorts of trees of different species, age and structure compete for light and soil resources on a number of replicate patches (15 patches of 1000 m^2 in the present study). Either plant functional types (PFTs) (Sitch et al., 2003) or species (Hickler et al., 2004; Koca et al., 2006) may be simulated.

Typical model output consists of leaf area index (LAI), net primary production (NPP), biomass, tree density, carbon fluxes and runoff. Values are averaged over the replicate patches to give stand averages of the relevant variables.

Using a very similar model set-up to that used here, Miller et al. (2008) showed that LPJ-GUESS could successfully model the Holocene dynamics of the main tree species at four sites in Fennoscandia where vegetation reconstructions using pollen accumulation rate data were possible.

Species description

In Table 2.1, we list the seventeen tree and shrub species, and the single grass taxon, used in the model, as well as their plant characteristics and bioclimatic limits. Further changes to the model parameters described by Smith et al. (2001), Hickler et al. (2004) and Miller et al. (2008), are listed Tables A.1 and A.2.

Species	Description	GDD_{5min} (°C d)	$T_{c_{min}}$ (°C)	$T_{c_{max}}$ (°C)	DT
Abies alba	T,Te,NE,St	1800	-4.5	-1	0
Alnus incana	T,Bo,BS,Ist	500	-30	-2.5	0
Betula pendula	T,Te,BS,Si	700	-30	7	0
Betula pubescens	T,Bo,BS,Si	300	-	6	0
Carpinus betula	T,Te,BS,Ist	1100	-8	5	1
Corylus avellana	T,Te,BS,Ist	700	-13	10	1
Fagus sylvatica	T,Te,BS,St	1300	-3.5	6	0
Fraxinus excelsior	T,Te,BS,Ist	1100	-10	6	0
Picea abies	T,Bo,NE,St	650	-30	-1.5	0
Pinus sylvestris	T,Bo,NE,Ist	450	-30	-1.0	1
Pinus halepensis	T,Te,NE,Ist	3000	3	9	1
Populus tremula	T,Te,BS,Si	500	-30	6	0
Quercus coccifera	S,Te,BE,Ist	3100	3.5	11	1
Quercus ilex	T,Te,BE,Ist	2000	0	10	1
Quercus robur	T,Te,BS,Ist	1100	-9	7	1
Tilia cordata	T,Te,BS,Ist	1100	-11	5	0
Ulmus glabra	T,Te,BS,Ist	850	-9.5	6	0
C3 grass	-,-,-,-	0	-	-	1

Table 2.1: Selected species with their characteristics and bioclimatic limits as specified in the model. The plant characteristics are: either trees (T) or shrubs (S), either boreal (Bo) or temperate (Te), either broadleaf summergreen (BS), broadleaf evergreen (BE) or needleleaf evergreen (NE), and either shade tolerant (St), intermediately shade tolerant (ISt) or shade intolerant (Si). The bioclimatic limits are: GDD_{5min} (minimum growing degree-day sum (5°C base)), $T_{c_{min}}$ (minimum temperature of the coldest month), $T_{c_{max}}$ (maximum temperature of the coldest month) and DT (drought tolerance).

The species are trees (T) or shrubs (S), boreal (Bo) or temperate (Te), broadleaf summergreen (BS), broadleaf evergreen (BE) or needleleaf evergreen (NE), and shade tolerant (St), intermediately shade tolerant (ISt) or shade intolerant (Si) (Smith et al., 2001). Trees and shrubs have different allometric relationships, and summergreen species require varying periods of chilling to induce budburst (Murray et al., 1989). The generic C3 grass PFT is intended to represent the numerous understorey species that are not considered in this paper, but nevertheless compete with trees for water and nutrients.

The maximum range limits of the tree species are defined in LPJ-GUESS by four key, species-specific bioclimatic constraints (Prentice et al., 1992; Sykes et al., 1996): GDD_{5min} (minimum growing degree-day sum (5°C base)), $T_{c_{min}}$ (minimum temperature of the coldest month), $T_{c_{max}}$ (maximum temperature of the coldest month) and DT (drought tolerance). Drought intolerant species ($DT = 0$) require an average growing season available water content of 30mm for establishment. The values in Table 2.1 were taken from the literature (Prentice and Helmisaari, 1991; Sykes et al., 1996), with minor adjustments prompted by comparison with European species distributions. The use of this minimal set of bioclimatic constraints, each of which represents a known or likely physiological limiting mechanism (Woodward, 1987; Miller et al., 2008), is more robust through time than simple correlations between climatic variables and species ranges.

For a species within its bioclimatic limits, cohort establishment and mortality are modelled yearly in LPJ-GUESS as stochastic processes within each replicate patch of the stand (Smith et al., 2001). Two additional stochastic processes are also considered in LPJ-GUESS. First, patch-destroying disturbances, representing destructive processes such as herbivory and storm damage, result in all vegetation in a patch being transferred to the patch's litter pool with a certain annual probability that is the inverse of the average disturbance interval of 100 years. Second, the yearly probability of a fire disturbance is modelled as in Thonicke et al. (2001).

The species listed in Table 2.1 are clearly a small subset of the full range of species seen in Holocene pollen diagrams. However, use of a restricted set was a necessary compromise. A larger species set would have increased the computational time required for model inversion. By choosing a restricted set containing a representative sample of

the diversity of vegetation and functional types seen in sub-Arctic Europe today, we expect to capture the main variability seen in the Holocene pollen records. Our choice was also restricted by the relatively small set of species with bioclimatic limits used by LPJ-GUESS that are known with any great degree of certainty.

Vegetation

From the different vegetation outputs of LPJ-GUESS: NPP, LAI, biomass, we summarise vegetation by using an average of simulated NPP over 30 years. This choice is driven by the need for maximum coherence between pollen samples and the summarises of the vegetation simulated at the same sites. Preliminary attempts to link pollen and these outputs convinced us that LAI and NPP are nearly equivalent and perform better than biomass which represents an accumulation of carbon mass in time. Thirty-year means for NPP correspond approximately to the accumulation period for pollen in modern samples.

We denote the simulated vegetation at the N modern sites as $V_{s=1:N}$, represented by the NPP averaged by species over 30 years. V_t is the mean of simulated vegetation, for the past, during the years $t-30$ to t. Note that all these elements are positive or null and that they represent absolute or "raw" production values.

2.2.2 Climate data

LPJ-GUESS is forced with chronologies of monthly precipitation, temperature and cloudiness. For each pollen site, we interpolated precipitation and temperature time series from the CRU TS 1.2 dataset (New et al., 2002). We used an ordinary kriging method with altitudinal gradient as an external drift (e.g. Cressie, 1991). For cloudiness we fitted a logit-linear regression between monthly cloudiness and both monthly precipitation and temperature per site.

The interpolated climate series are considered as a skeleton to which anomalies are

applied to determine the optimal fit between model outputs and pollen data. These anomalies will be referred to as parameters or climate parameters in the following sections because they are the climate quantities which are reconstructed. We denote by C_s for modern sites and C_t for core at time t, the 6-dimensional climate parameter vector: $C = (T_{\text{jan}}, T_{\text{jul}}, P_{\text{win}}, P_{\text{spr}}, P_{\text{sum}}, P_{\text{aut}})$. First parameters are absolute temperature anomalies (in °C) from January and July 20th century series. The precipitation parameters are relative anomalies (in %). Let $T_{(i,j)}$ be the original temperature of year i and month j, $T_{(.,j)}$ the 100-year mean temperature of month j, where $j = 1$ denotes January. Then the transformed temperature $\tilde{T}_{(i,j)}(C)$ is defined as a function of parameters T_{jan} and T_{jul}:

$$\tilde{T}_{(i,j)}(C) = T_{(i,j)} - T_{(.,j)} + (T_{(.,j)} - T_{(.,1)}) * \left(\frac{T_{\text{jul}} + T_{(.,7)} - (T_{\text{jan}} + T_{(.,1)})}{T_{(.,7)} - T_{(.,1)}} \right) + T_{\text{jan}} + T_{(.,1)}$$

This transformation modifies the monthly mean temperature signal ($T_{(.,j)}$) by scaling it to match new January and July specified temperatures T_{jan}+$T_{(.,1)}$ and T_{jul}+$T_{(.,7)}$. The interannual variability of the transformed series is exactly the same as in the original skeleton.

Precipitation parameters are seasonal percentages which are added to the original skeleton. Let $P_{(i,j)}$ be the original precipitation of year i and month j. Then the P_{win} anomaly is applied by multiplying each winter month (January, February and March) by $(1 + P_{\text{win}}/100)$ to obtain $\tilde{P}(C)$ modified precipitation

$$\tilde{P}_{(i,j=1:3)}(C) = P_{(i,j=1:3)} * (1 + P_{\text{win}}/100)$$

Here the positivity constraint of precipitation is respected, but the interannual variability of the original chronologies is modified.

Once the modified $\tilde{T}(C)$ and $\tilde{P}(C)$ have been created, a modified sunshine $\tilde{S}$ is computed by regression using $\tilde{T}(C)$ and $\tilde{P}(C)$ as regressor variables.

2.2.3 Pollen data

Pollen surface samples

The pollen surface sample database has been compiled by Bordon (2008) from data taken from Bottema (1974), Brugiapaglia (1996), Peyron et al. (1998) and Sanchez Goni and Hannon (1999).

The database was initially composed of 1512 different modern sites covering Europe and Morocco with more than 150 different pollen taxa. A subselection of taxa was made to correspond to the output of the model. We computed 14 groups by summing taxa corresponding to each of the following 14 arboreal taxa: Abies, Alnus, Betula, Carpinus, Corylus, Fagus, Fraxinus, Picea, Pinus, Quercus Evergreen, Quercus Deciduous, Tilia, Ulmus and Populus. A "grasses and shrubs" (GrSh) group was made by summing all non-arboreal and non-aquatic taxa. This selection preserves a maximum of coherence with the 18 species (or groups of species) defined in the version of LPJ-GUESS that we use. See Table 3.1 for the correspondence between pollen groups and vegetation model species.

We first filtered the sites by removing all non-terrestrial sample sites due to spurious coordinates or offshore core tops (eg. in Danube estuaries). Offshore pollen samples are representative of pollen production of a whole watershed and are not coherent with other more local records. As a second selection criterion we removed all samples where the taxa subselection resulted in a loss of more than 25% of the original pollen count. These cases occurred when more than 25% of the sample consisted of arboreal taxa other than those simulated by LPJ-GUESS. We consider that the removal of such a large part of the original pollen spectra would result in a distorted image of the surrounding vegetation.

Since most of the pollen samples where available as percentages (more than 70%) we converted counted samples to percentages.

The final dataset is a matrix containing $N = 1209$ modern sites (rows) and 15 taxa

i	pollen type: y_i	vegetation species: $v_{j(i)}$
1	Abies	Abi_alb
2	Alnus	Aln_inc
3	Betula	Bet_pen + Bet_pub
4	Carpinus	Car_bet
5	Corylus	Cor_ave
6	Fagus	Fag_syl
7	Fraxinus	Fra_exc
8	Picea	Pic_abi
9	Pinus	Pin_syl + Pin_hal
10	QuercusE	Que_coc + Que_ile
11	QuercusD	Que_rob
12	Tilia	Til_cor
13	Ulmus	Ulm_gla
14	Populus	Pop_tre
15	GrSh	C3_gr

Table 2.2: Correspondence table between pollen types and species defined in LPJ-GUESS.

per site (columns). The spatial distribution of this dataset is shown in Figure 2.1. A modern pollen sample $Y_s = (Y_{s,1}, Y_{s,2}, \ldots, Y_{s,15})$ is a 15-dimentional vector representing the pollen proportion per group, where $\Sigma_{j=1}^{15} Y_{s,j} = 1$.

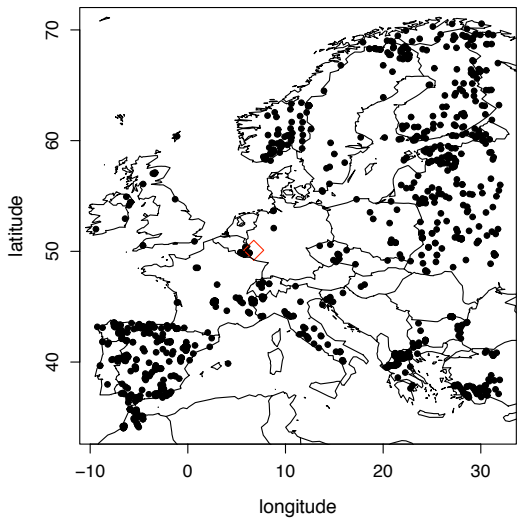

Figure 2.1: Distribution of the 1209 modern pollen samples (black dots) and location of the Meerfelder Maar site (square).

Meerfelder Maar sediment core

The sediment core (Litt et al., 2009) was taken from the lake Meerfelder Maar (50.1°N,6.75°E, see Figure 2.1) located within the Westeifel Volcanic Field less than 10 km apart. The uppermost 180 cm of the core, corresponding to approximately the last 1.6 cal ky BP (calendar kilo-years Before Present; refers to the number of years before 1950), are not continuously varved and were dated using two AMS ^{14}C dates and extrapolated sedimentation rates based on varve data. The other part of the core is varved and the endpoint of the core has been linked to a calendar-year chronology by using a tephra dated at 11 cal ky BP. In total 406 samples have been collected and analysed from this core. The number of pollen grains counted in each sample is between 500 and 1000 and we transform it to percentages to agree with modern data.

The pollen diagram is presented Figure 2.2. For the interpretation of climate reconstruction results we divide the 11 to 0 cal ky BP chronology in four periods. A more comprehensive discussion can be found in Litt et al. (2009). The earliest period (11 to 10 cal ky BP) corresponds to the end of the glacial period and is characterised by a rapid decrease of Grass-Shrub group, Pinus and Betula. This decline is matched by marked increases in Corylus. The 10 to 6.3 cal ky BP phase shows a Corylus decrease and the arrival of Ulmus, followed by Tilia, Fraxinus and Alnus and, finally, Fagus. The 6.3 to 3.7 phase is a very stable one with high levels of Alnus, but less Ulmus and Tilia than during the previous period. The last period 3.7 to 0 cal ky BP is characterized by a number of changes, due to increasing anthropogenic pressure and climate changes. There is a global but non-monotonic increase of the Grass-Shrub group and a non-monotonic decrease of Alnus, Ulmus and Tilia.

Sediment core samples are denoted $Y_{t=t_1:t_n}$. The core contains $n = 406$ samples from $t_1 = 10988$ cal yr BP to $t_{406} = 0$ cal yr BP. As for the modern samples, each sample Y_t is a 15-dimentional vector representing the pollen proportion per pollen group.

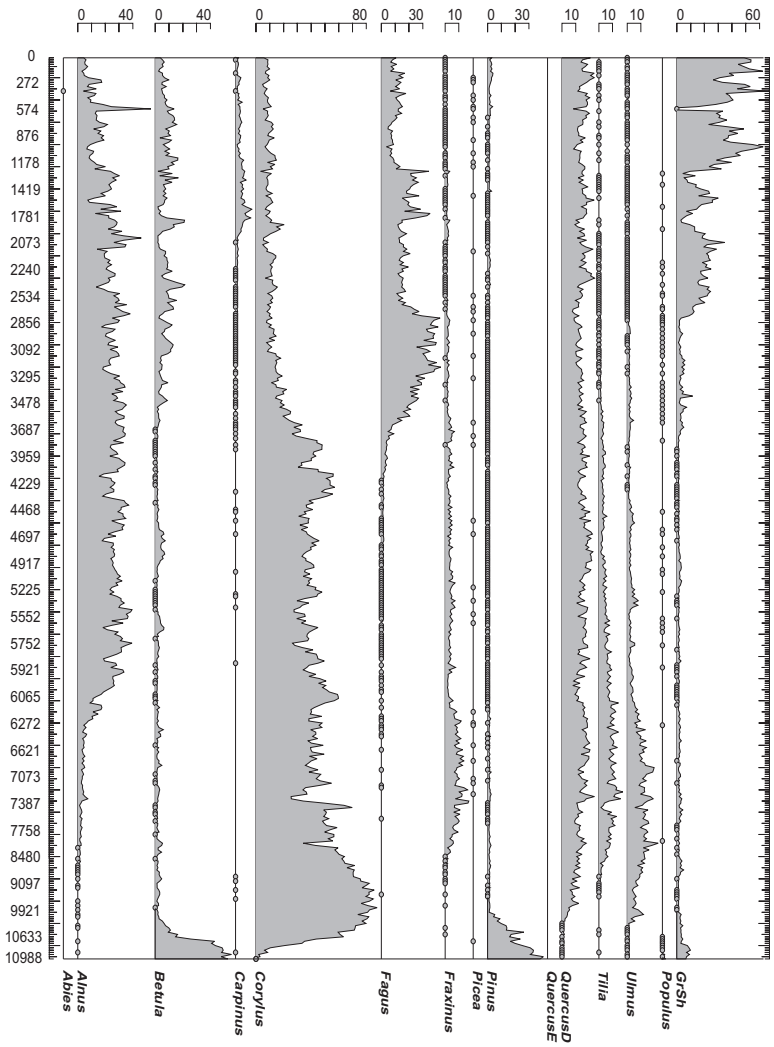

Figure 2.2: Pollen diagram from the Meerfelder Maar sediment core. (x-axis) The fifteen pollen groups defined in Table 3.1 are given as percentages (over the fifteen groups) and (y-axis) the age is in calendar years Before Present (BP, refers to before 1950) from 0 (top) to 10988 (bottom). The total number of pollen grain counted per sample range between 500 and 1000. The uppermost 180 cm of the core, corresponding to approximately the last 1.6 cal ky BP, are not continuously varved and were dated using two AMS ^{14}C dates and extrapolated sedimentation rates based on varve data. The other part of the core is varved and the endpoint of the core has been linked to a calendar-year chronology by using a tephra dated at 11 cal ky BP (Litt et al., 2009). Therefore the uncertainties associated to the chronology are not constant along the core and very hard to quantify. For the period of time between 11 cal ky BP and around 1.6 cal ky BP, varved sediments imply that there is no uncertainty *between* sample dates; but this is a floating chronology implying a constant uncertainty for the overall time-period. Uncertainties after 1.6 cal ky BP and for the whole floating chronology had a magnitude of around 100 yr but were corrected by stratigraphic alignment with the well dated core of the Holzmaar Maar, which reduce them.

2.2.4 A statistical model to link climate, vegetation and pollen

We build a statistical model which embeds the vegetation model and describes the relations between the variables climate C, vegetation V and pollen Y. This is designed for pollen samples taken from a single sediment core. Each date $t = t_1 : t_n$ is considered known without uncertainty. This statistical model is *hierarchical Bayesian*.

Hierarchical means that it is based on a conditional "split" of the model. For comparison, a Transfer Function (TF) models a direct link between climate and pollen using the conditional distribution of pollen given climate $p(Y|C)$. In this work, we model $p(Y|C)$ hierarchically by specifying a distribution of the vegetation conditional on climate $p(V|C)$ and a distribution of the pollen conditional on vegetation $p(Y|V)$.

Bayesian theory is a framework for inference (Young and Smith, 2005). In the context of this applied work we use the main concepts of "prior" and "posterior". The prior is the information, summarised under the form of a distribution, which is available prior to the data analysis. For example we will use a prior on climate at time t, $p(C_t)$. This is the information on ancient (time t) climate we have before running the inversion, and may be estimated by climate reconstructions already available before inversion. After the choice of a prior on climate $p(C)$ and a hierarchical model $p(V|C).p(Y|V)$, the Bayesian inference consists in obtaining the posterior distribution of climate and vegetation given pollen $p(C, V|Y)$. The Bayes theorem gives the link between the prior, the structure and the posterior: $p(C, V|Y) = p(C).p(V|C).p(Y|V)/p(Y)$.

The structure of the hierarchical model is illustrated on the graphic Figure 2.3. It is based on the basic elements $p(C)$, $p(V|C)$ and $p(Y|V)$ described above, and each individual part is described in more details below. The choice of prior distribution $p(C)$ is discussed below in each reconstruction exercise. In the next section we define and calibrate the distribution of pollen given vegetation $p(Y|V)$. Section 2.2.4 completes the definition by describing $p(V|C)$ as the vegetation model.

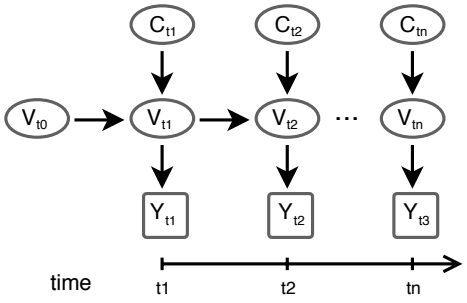

Figure 2.3: Graphical representation of the hierarchical model for reconstruction. Considered variables are: C the climate, V the vegetation simulated by LPJ-GUESS and Y the pollen data. Subscripts indicate a sample date (t) for the analysed core. Known variables are in a square and variable to be reconstructed in a circle. Arrows represent the conditioning between variables. The times t_1 to t_n are the core point dates. t_1 is the oldest age sampled in the core, t_n the most recent. Pollen data are known for each sample of the core points. Following the arrows we see that vegetation at time t_2 (V_{t_2}) depends on climate at time t_2 (C_{t_2}) and vegetation at time t_1 (V_{t_1}). This is modelled using LPJ-GUESS forced with climate C_{t_2}, starting from V_{t_1} and run during $t_2 - t_1$ years. Pollen at time t_2 (Y_{t_2}) depends only on V_{t_2} through a $p(Y|V)$ distribution.

Calibration of the pollen/vegetation distribution

A key element of the inversion is the relationship between simulated vegetation and pollen data $p(Y|V)$. In statistics this is called the pollen likelihood and can be compared to a transfer function between vegetation and pollen. We model it using non-parametric kernel smoothed surfaces. These surfaces are calibrated using the modern pollen dataset and modern simulations of the vegetation.

The distribution $p(Y|V)$ models the relationships between 15 pollen proportions Y and the simulated NPP of 18 species, V. Thus its dimension is $15 + 18 = 33$ and it contains information about $15 * 18 = 270$ variable crossings. We reduce this dimension as follows:

$$
\begin{aligned}
p(Y|V) &\approx \prod_{i=1}^{15} p_i(Y_i|V) && (2.1)\\
&\approx \prod_{i=1}^{15} p_i(Y_i|V_{j(i)}) && (2.2)\\
&= \prod_{i=1}^{14} q_i(Z_i|V_{j(i)}) && (2.3)\\
&\approx \prod_{i=1}^{14} \tilde{q}_i(Z_i|V_{j(i)}) && (2.4)
\end{aligned}
$$

subject to the following assumptions:

- (2.1) conditional on vegetation, all pollen abundances Y_i, Y_j for $i \neq j$ are independent. This is acceptable since, with a pollen time resolution of 20 to 30 years, given the vegetation, pollen production of one group can be considered independent of the production of all other pollen group.

- (2.2) all information about Y_i is carried by only one $V_{j(i)}$ species of vegetation. The subscript $j(i)$ refers to the j^{th} vegetation species corresponding (i.e. as a function of) to the i^{th} pollen group. This is acceptable if there is a good agreement between pollen groups and the species simulated by LPJ-GUESS. These correspondences are specified in Table 3.1.

- (2.3) using the variable transformation $Z = alr(Y)$ (Aitchison, 1982) there always exists a relationship, called q_i, between vegetation V and the transformed pollen variable Z. Pollen variables are probabilities implying that their sum is $\sum_{i=1}^{15} Y_i = 1$. This constraint reduces the dimension to 14 since the fifteenth proportion is defined by $1 - \Sigma_{i=1}^{15} Y_i$. First we set $Y_{i,j} = 10^{-4}$ for all (i, j) where $Y_{i,j} = 0$. Then we apply the Aitchison (1982) transformation $Z = alr(Y)$ defined as:

  $$Z_{i=1:14} = \log(Y_i/Y_{15})$$

  This reduces the dimension to 14 and lets us model the unconstrained variables Z_i.

- (2.4) each q_i is correctly approximated by a $\tilde{q}_i$ obtained by kernel smoothing (e.g. Loader, 1999).

In practice, the fourteen surfaces q_i are fitted on the modern dataset $(C, V)_{s=1:N}$. For this purpose we use a gaussian kernel whose parameters are fitted by cross validation (Loader, 1999). Figure 2.4 shows the obtained surfaces.

LPJ-GUESS is a vegetation/climate distribution

Let $p(V_{t_j}, C_{t_j}|V_{t_i})$ be the distribution for the temporal transition of climate and vegetation from time t_i to time t_j where t_i and t_j are dates for consecutive samples of the core

$$p(V_{t_j}, C_{t_j}|V_{t_i}) = p_{\mathrm{LPJ}}(V_{t_j}|V_{t_i}, C_{t_j}).p(C_{t_j}) \tag{2.5}$$

with $p(C_{t_j})$ the prior climate distribution at t_j. $p_{\mathrm{LPJ}}(V_{t_j}|V_{t_i}, C_{t_j})$ is (defined by) the randomness of V_{t_j} when we run LPJ-GUESS for $t_j - t_i$ years starting from vegetation V_{t_i}, with climate C_{t_j}.

Vegetation simulated by LPJ-GUESS is stochastic in the sense that several runs of the vegetation model with the same forcings give slightly different vegetation values. We therefore consider vegetation as a random variable and LPJ-GUESS as a distribution: $p_{\mathrm{LPJ}}(V_{t_j}|V_{t_i}, C_{t_j})$. This distribution simulate the variable V_{t_j} and has parameters (V_{t_i}, C_{t_j}) but also the climate chronologies and numerous forcings like soil, CO_2 etc. We can sample from this distribution by running the model, but since it is a complex computer code we cannot compute its probability value for any given set of variables and parameters. This represents a major change from deterministic vegetation models such as BIOME3 (Haxeltine and Prentice, 1996) or the LPJ-DGVM (Sitch et al., 2003).

The temporal link of the hierarchical model is given by LPJ-GUESS and arises from the later definition. To simulate vegetation at time t_j younger than (after) t_i, where t_i and t_j are the dates of consecutive core samples, the vegetation model starts with V_{t_i}

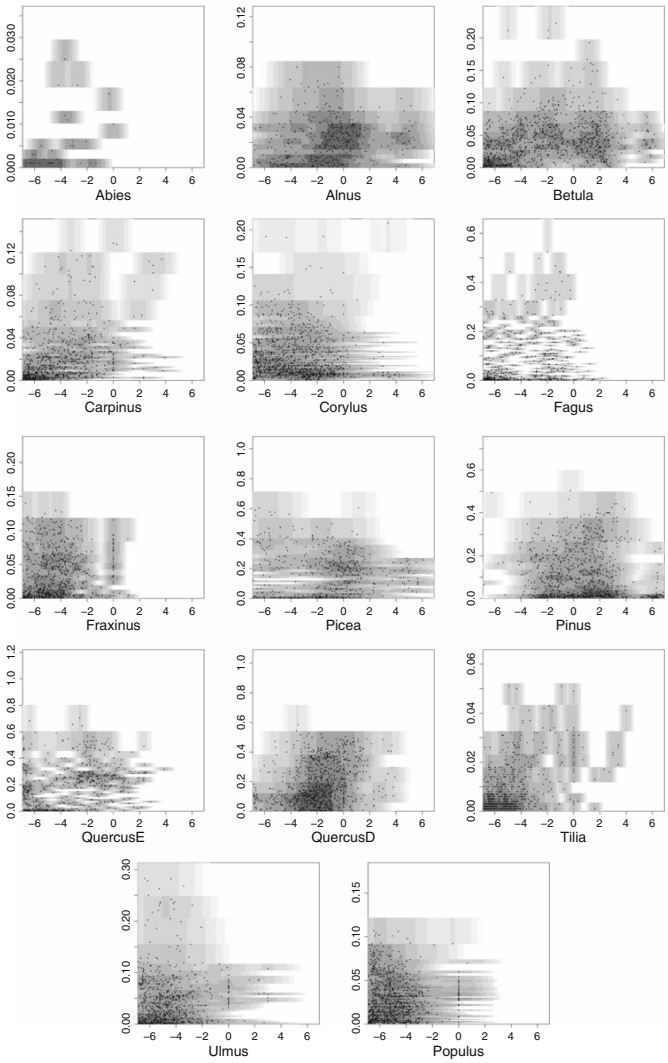

Figure 2.4: Kernel smoothed surfaces for the 14 groups. The joint smoothings $q_i(Z_i, V_{j(i)})$ are derived from the conditional smoothings used for inference: $q_i(Z_i|V_{j(i)})$. Graphics are in the same order as in Table 3.1. Each plot presents (x-axis) a pollen group transformed following Aitchison (1982) (Z_i without unit) versus (y-axis) its corresponding simulated annual net primary production (NPP, in kg carbon m^{-2} $year^{-1}$). The dots are the modern data and the shading shows the density obtained by kernel smoothing (darker means higher density).

and runs for $t_j - t_i$ years. If $t_j - t_i$ is short, the vegetation simulated at t_j is strongly forced by vegetation V_{t_i} and then, implicitly, by climate C_{t_i}. This constraint gives a time-coherence to the reconstructed vegetation.

2.2.5 Inference using a particle filter algorithm

In the Bayesian context, reconstruction of climate and vegetation involves the computation of the joint posterior distribution $p(C_{t_1:t_n}, V_{t_1:t_n}|Y_{t_1:t_n})$. This represents the distribution of climate and vegetation "histories" from time t_1 to t_n knowing all $Y_{t_1:t_n}$ pollen data.

Particle filters provide a reconstruction based on Importance Sampling (IS) which is sequential, i.e. done sample after sample. The sequential aspect solves the curse of dimensionality because it slices the climate space. A simple explanation follows: at time t_j the algorithm has a reconstruction obtained for the preceding point t_i. A set of 1000 possible climates $C_{t_j}^{(l)}$ is proposed from the prior $p(C_{t_j})$. LPJ-GUESS is then run for each climates starting from the reconstructed vegetation at t_i and for the years t_i to t_j. It produces couples of climate and associated vegetation $(C_{t_j}, V_{t_j})^{(l)}$ that we call "particles". In this set of particles, a selection is done by comparison of each vegetation simulated and the pollen Y_{t_j}. This selection consists in computing $\omega^{(l)}$ equal to the likelihood $p(Y|V)$ of the pollen Y_{t_j} for each simulated vegetation $V_{t_j}^{(l)}$. High $\omega^{(l)}$ score means that the couple $(C_{t_j}, V_{t_j})^{(l)}$ is highly probable and null scores means that the couple is not coherent.

A full comprehensive description of the algorithm is given appendix A.2. We just give here a summary of the algorithm.

1. INITIALISATION :

   - Generate N_p couples $(V_{t_1}, C_{t_1})^{(l=1:N_p)}$, by sampling $C_{t_1}^{(l)}$ from $p(C_{t_1})$ and running LPJ-GUESS until equilibrium is reached with climate $C_{t_1}^{(l)}$ to obtain $V_{t_1}^{(l)}$,

- Compute for each particle $(V_{t_1}, C_{t_1})^{(l)}$ the weight $\omega_{t_1}^{(l)} = p(Y_{t_1}|V_{t_1}^{(l)})$ using the kernel smoothed surfaces,
- Compute each normalised weights $\tilde{\omega}_{t_1}^{(l)} = \omega_{t_1}^{(l)} / \Sigma_{k=1}^{N_p} \omega_{t_1}^{(k)}$

2. RESAMPLING

- Compute the criterion $\text{ESS}_t = \left(\Sigma_{l=1}^{N_p} \left(\tilde{\omega}_t^{(l)} \right)^2 \right)^{-1}$
- If $\text{ESS}_t < N_p/2$ randomly sample the particles by residual resampling and set all weights $\tilde{\omega}_t^{(l=1:N_p)} = 1/N_p$.

3. SAMPLING

- For current time t_j immediately consecutive to t_i,
- Sample N_p particles $(V_{t_j}, C_{t_j})^{(l=1:N_p)}$ by sampling $C_{t_j}^{(l)}$ from $p(C_{t_j})$ and running LPJ-GUESS starting from V_{t_i}, for t_j to t_i years with climate C_{t_j} to obtain $V_{t_j}^{(l)}$,
- Compute for each particle $(V_{t_j}, C_{t_j})^{(l)}$ the weights $\omega_{t_j}^{(l)} = \tilde{\omega}_{t_i}^{(l)}.p(Y_{t_j}|V_{t_j}^{(l)})$ using the kernel smoothed surfaces,
- Compute each normalised weights $\tilde{\omega}_{t_j}^{(l)} = \omega_{t_j}^{(l)} / \Sigma_{k=1}^{N_p} \omega_{t_j}^{(k)}$
- If $t_j < t_n$, then set $t_j = t_k$ (k=j+1) and go to the RESAMPLING step ***else stop***

2.3 Results

2.3.1 Validation using modern pollen samples

In this section our goal is to validate the statistical framework, and not the vegetation model. Since the climate-pollen relationship may be locally biased at any given site, due to local errors induced by pollen production changes, non-homogeneous transport, different accumulation processes, etc, the method must be tested at a series of sites.

We have therefore reconstructed modern climate at 30 sites randomly chosen from the modern pollen dataset.

The validation at each site $s = 1 : 30$ was performed as follows:

- repeat for each particle l=1:1000
  - sample $C_s^{(l)}$, one 6-dimensional climate parameter following the prior defined in the next section,
  - Spin-up LPJ-GUESS for 500 years with repeated 1901-1930 monthly climate to which is added $C_s^{(l)}$ using the 1901 value of CO_2 atmospheric concentration of 296.3 ppmv,
  - simulate the 1901-2000 vegetation using 1901-2000 monthly climate to which is added $C_s^{(l)}$ under evolving CO_2 atmospheric concentration as obtained from the Carbon Cycle Model Linkage Project (McGuire et al., 2001).
  - retain $V_s^{(l)}$, the mean of NPP over the years 1961-1990,
  - weight the couple $(C_s^{(l)}, V_s^{(l)})$ by $\omega_s^{(l)} = p(Y_s|V_s^{(l)})$.

Definition of climate prior

As each modern site represents a single point in time, priors were chosen based on the CRU 1.2 (New et al., 2002) gridded set of climatological data for the European continent. For each validation site $s = 1 : 30$, the climate prior

$$\begin{aligned} p(C_s) &= p(T_{\text{jan}}, T_{\text{jul}}).p(P_{\text{win}}).p(P_{\text{spr}}).p(P_{\text{win}}).p(P_{\text{aut}}) \\ &= N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{jan} & C_{jan,jul} \\ C_{jan,jul} & V_{jul} \end{pmatrix}\right) . \left(N_{t=0}(0, \sigma^2_{Prec})\right)^4 \end{aligned}$$

is composed of a bivariate Gaussian distribution for T_{jan} and T_{jul} and 4 times the same 0-truncated independent Gaussian distribution for each precipitation parameter. Each distribution is centred on 0, the null anomaly equal to expected climate. Variance and covariance parameters of the bivariate Gaussian law are derived from the CRU TS 1.2

(New et al., 2002) means for months January ($V_{\text{jan}} = 6.6^2$) and July ($V_{\text{jul}} = 4.1^2$). The covariance ($C_{jan,jul} = 18.67$) or correlation ($\rho = 0.69$) represent the modern European seasonal link between these two months. In doing so we allow temperature parameters for each site to be distributed over the whole modern European temperature set. The standard deviation for the precipitation parameter σ^2_{Prec} was arbitrarily chosen as 35 (in %) giving a probability of 0.005 to exceed an 100% precipitation increase.

Validation Results

For each site s we obtain a set of 1000 tested climates (the particles) $C_s^{(l)}$ associated to weights $\omega_s^{(l)}$. At each site, we summarise, the set of weighted climates by computing quantiles $q_{0.025}$, $q_{0.5}$ (median) and $q_{0.975}$. For visual representation we smooth the particles and obtain graphics showing the distribution of tested climates, for example Figure 2.5 obtained for a Spanish site (41.39°N, 0.11°W). For the validation, however, we are interested in the global result obtained for the whole set of 30 climate reconstructions. Figure 2.6 presents the observed January and July temperatures versus their reconstructions. Table 2.3 summarises the reconstruction results.

	Expected $q_{.5}$	Prior $q_{.975} - q_{.025}$	Posterior $q_{.5}$	Posterior $q_{.975} - q_{.025}$
T_{jan} (°C)	0	25.9	0.85	16.8
T_{jul} (°C)	0	16.1	0.82	12.3
P_{win} (%)	0	137	−1.6	148
P_{spr} (%)	0	137	−0.2	142
P_{sum} (%)	0	137	0.7	143
P_{aut} (%)	0	137	−2.6	138

Table 2.3: Means of the results obtained for 30 European points for the validation using present-day pollen samples. Rows are the 6 reconstructed climate parameters. Columns are: (1) the expected median (0.5 quantile), (2) the prior 0.95 bilateral interval width (equal to 0.975 quantile minus 0.025 quantile), (3) the posterior median and (4) the posterior 0.95 bilateral interval width.

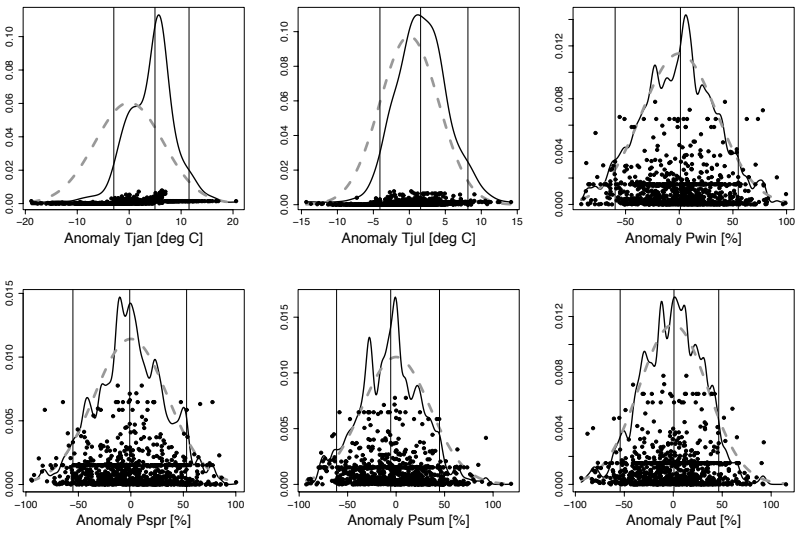

Figure 2.5: Prior (grey dashed lines) and posterior (black lines) smoothed distributions of the 6 climate anomalies: $(T_{\text{jan}}, T_{\text{jul}}, P_{\text{win}}, P_{\text{spr}}, P_{\text{sum}}, P_{\text{aut}})$. The "particles" of climate proposed by the particle filter are the black dots. Black thin vertical lines show the 2.5%, median and 97.5% quantiles of each posterior distribution. This is an example for a dry Mediterranean site from Spain (41.39°N, 0.11°W).

The mean discrepancies between posterior medians and expected values of the 6 reconstructed parameters are negligible by comparison with interval widths (see Table 2.3). All the confidence intervals of the reconstructed temperatures contain the observed values (see Figure 2.6). Thus, the method seems to be unbiased, at least at the continental level. Further, the temperature posteriors distributions are narrower than priors (see Table 2.3 and Figure 2.6). In other words, the inversion process is able to constrain both temperature variables from the specified prior.

Precipitation posteriors are not narrower than their priors. This shows that the inversion process is unable to constrain four precipitation variables (at a time) further than what has been specified as prior. That the vegetation model does not show a precipitation constraint is surprising. It may be assumed that in a non water-stressed region such as a European temperate forest an increase or a moderate decrease in pre-

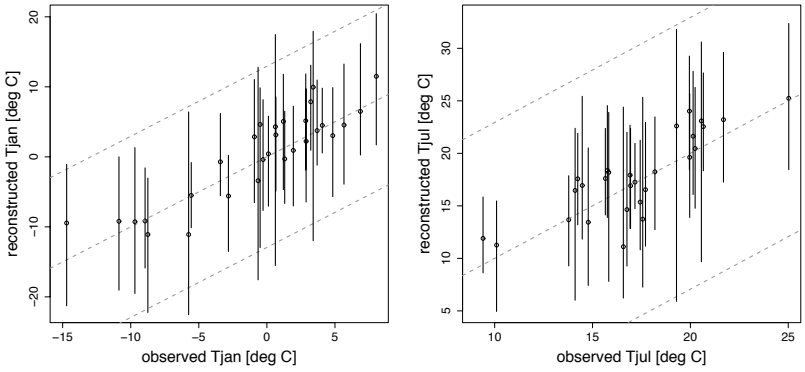

Figure 2.6: Observed values versus reconstructed (posterior) means and 95% confidence intervals for January and July temperatures in Celsius degrees for the 30 validation sites. The grey dashed lines represent the prior mean and quantiles. Note that prior mean is the expected (equal to observed) value. The black lines give the posterior interval range and the point is the reconstructed mean. (left) January temperature and (right) July temperature.

cipitation would not change vegetation composition dramatically. However, in a highly water-stressed region, such as the Mediterranean dry region, water availability is one of the main vegetation drivers. We have therefore further investigated the reasons for this lack of constraint and found that it is due to a combination of the vegetation model and the prior chosen for precipitation.

Vegetation model: We performed a sensitivity analysis using the inversion algorithm. A second validation was performed at 20 modern Mediterranean dry points. These points were randomly selected from the set of points below latitude 42°N with less than 600mm of rain per year, and with less than 70% of tree taxa pollen in their samples. For each climate proposed by the algorithm, $C_s^{(l)}$, we computed at a Drought Stress Indicator (DSI), $R_s^{(l)}$ based on quantities calculated by the vegetation model. This DSI is defined as the ratio between annual mean actual evapotranspiration (AET) and annual mean potential evapotranspiration (PET, e.g. Sykes et al., 1996). We found that, for

a varying climate $C_s^{(l)}$, the DSI varies in $[0.10; 0.50]$ which corresponds to a xerophytic vegetation (Prentice et al., 1992). Thus, the vegetation simulated for these dry sites agrees with pollen data. This indicates two things. First, the link vegetation/pollen $p(Y|V)$ performs well since we obtain a coherent match between simulated vegetation and sampled pollen. Second, the vegetation model seems to underestimate DSI variation as a function of precipitation change since the most humid $C_s^{(l)}$ (proposed climates) should result in higher DSI than 0.5 and non-xerophytic vegetation.

The priors for precipitation: When specifying independent priors for the precipitation per season we implicitly specify a small relative range for the annual precipitation. This is due to compensation between seasons. In other words, to sample extreme annual precipitation it is necessary to sample extreme precipitation for the majority of seasons, which has a very small probability of occurring. A simple way to scan a large range for annual precipitation is to define a single annual precipitation parameter, but this would fix seasonality. For the Meerfelder Maar reconstruction we chose to let seasonality vary by using four precipitation parameters.

2.3.2 Temporal model inversion on the Meerfelder Maar pollen sediment core

Prior definition

In an application to samples from a sediment core, the climate priors have to be elicited (obtained from expert's knowledge) or obtained using other data than those used for reconstruction. For example one can use the Modern Analogue Method (MAT, Guiot and De Vernal, 2007) and data for other sites obtained from the European Pollen Database (EPD, `http://www.europeanpollendatabase.net`) to build a prior distribution of climate for the studied European site during the Holocene. For Meerfelder Maar, since our goal is to present the method, we used an empirical prior based on MAT reconstruction of the climate using the same core. This allows us to assess how much the inversion

approach modifies the standard MAT estimates.

We applied the MAT with the Meerfelder Maar sediment core to reconstruct the six climate anomalies ($T_{\text{jan}}, T_{\text{jul}}, P_{\text{win}}, P_{\text{spr}}, P_{\text{sum}}, P_{\text{aut}}$). For the analogue dataset, we used the modern pollen data and climate described above. We used the classic chord distance between pollen samples. Following Guiot and De Vernal (2007) we computed by cross validation a discriminant distance for the analogue selection and we selected a maximum number of 7 analogues if this distance is not reached. We used the means reconstructed by MAT as prior means. We used the same temperature standard deviations as in the modern validation exercise ($sd(T_{jan}) = 6.6$ and $sd(T_{jul}) = 4.1$). The correlation between them was reduced (to 0.5), however, to relax the constraint on temperature seasonality during the Holocene. Standard deviation of the precipitation is 35 (in %).

Reconstruction results

For atmospheric CO_2 input, we used a composite record composed of the ice core record from Indermuhle et al. (1999) for the period 11 cal ky BP to cal 990 cal yr BP and the one from Siegenthaler et al. (2005) for the period 990 to 0 cal yr BP.

Posterior reconstructions of January and July temperatures and annual precipitation are presented Figure 2.7. The main events appearing in the pollen diagram (Figure 2.2), compared to the climate reconstruction in Figure 2.7 are:

- from about 10.6 cal ky BP, Betula and Pinus are replaced first by Corylus: this is the major event of the sequence indicating a warming of more than 10°C in winter and 5°C in summer and a precipitation increase of more than 500 mm/yr.
- After 10 cal ky BP, first Ulmus and Quercus deciduous, second Tilia and finally Fraxinus appear according to the classical succession in Europe; this does not translate any significant change in our climate reconstruction.
- Just before 6 cal ky BP, Alnus becomes dominant over Fraxinus, Ulmus and Tilia: this seems to indicate a slight increase of temperature (a few deegres Celsius) and of precipitation (100-200 mm/yr).

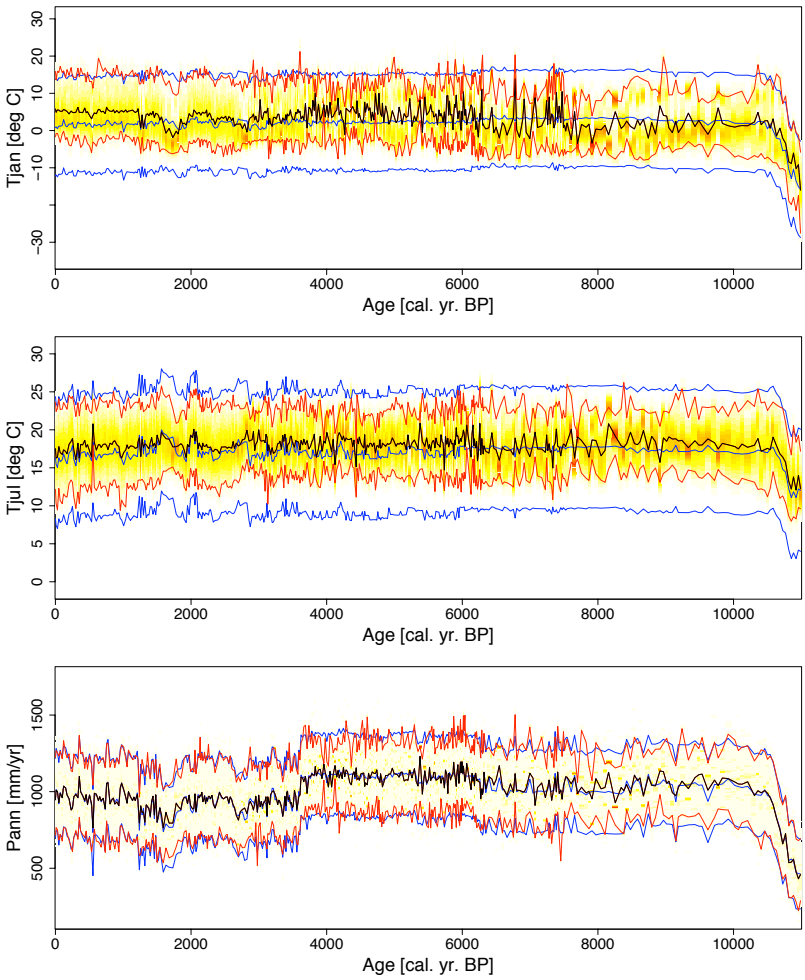

Figure 2.7: Climate reconstruction for Meerfelder Maar sequence during the Holocene using prior based on MAT estimations. Prior (blue lines): mean and 95% confidence interval. Posterior (red lines) mean (highlighted with black) and 95% confidence interval. The background colour (yellow to red) shows the posterior smoothed (low to high) density. For the three plots, x-axis shows the age in cal yr BP, y-axis shows: (top) January reconstructed temperature in ^{o}C, (center) July reconstructed temperature in ^{o}C and (bottom) annual reconstructed precipitation in mm/yr. Values recorded for the period 1961-1990 at the nearest meteorological station show a January mean temperature of -0.3^{o}C and 16.3^{o}C for July. Mean annual precipitation is 908 mm/yr (Litt et al., 2009).

- The second major event is at about 3.5 cal ky BP, Corylus is replaced by Fagus and secondly by Betula; this seems to indicate a slight decrease of January temperature (a few deegres Celsius) and an important precipitation decrease of about 300 mm/yr.

- At about 2.5 cal ky BP, grass and shrub group (GrSh) becomes important and is dominant after 1 cal ky BP. The anthropogenic deforestation should be translated by a reconstructed increase of drought, but there is nothing of such here. This seems to indicate that the method is robust against human disturbance.

This comparison of both figures show that the climate variations, as reconstructed by model inversion, are coherent with the pollen curves and seems to be robust against anthropogenic disturbance. This is a major argument in favor of these results.

The confidence intervals of January and July reconstructed temperatures have widths ranging from 10 to 20°C. This is coherent with, in the Gaussian case, a standard deviation of 2.5 to 5°C. These large posterior intervals are partly due to the large prior intervals. We have chosen to show means and quantiles here, but since the posterior is highly non-Gaussian the median or mode(s) of the posterior distribution may be preferred. A discussion of the results obtained follows:

In the earliest part of the sequence, 11 to 10.5 cal ky BP, corresponding to the transition to the Holocene period, there is a good agreement between our reconstructions and the MAT estimations.

During the period 10.5 to 7.5 the T_{jan} reconstruction shows two sets of possible values (high probability in dark red). In the first set, temperatures of around 10°C are reconstructed, which is higher than the MAT means. In the second set, the temperatures are lower at around -2°C. We note that these sets are non-continuous; the most probable climate jumps between paths. This results from the lack of an explicit climate link between samples in our statistical model.

At 7.5 cal ky BP the T_{jan} confidence interval becomes tighter. This is likely to be an

artifact resulting from the spike in Corylus pollen at this date. This causes a distortion in the diagram and therefore in the model as it tries to reproduce this abrupt peak.

The period 6.3 to 3.7 cal ky BP is a stable period in the pollen diagram, however, both January and July reconstructed temperatures show high frequency variability. This is again due to the absence of climate correlation in our model and may be further associated to an overfitting of the $p(Y|V)$ distribution. This would cause the vegetation model to reproduce non-significant changes in pollen samples, and give these a climatic interpretation. As there is no constraint in climate change through time, we cannot limit the reconstructed climate by reference to the value obtained for the previous sample. This results in non-significant fluctuations in the reconstruction.

The last 3.7 to 0 cal ky BP period is characterised by more marked changes in pollen composition. Here, however, the high variability in the reconstruction disappears, mean variability is more coherent in time and varies around the MAT estimates. This suggests that the changes in pollen composition during this period are sufficiently large to adequately constrain the reconstructed climate in adjacent samples, and result in a smoother curve.

The prior and posterior distributions of annual precipitations are similar except for a short period around 8 cal ky BP. This is unsurprising, as the validation exercise indicated that there is little or no constraint on precipitation. However, as Litt et al. (2009) found a quite different pattern for precipitation with a value close to 600 mm/yr at the top of the core instead of our value of around 1000 mm/yr, we devised a second test to check the precipitation constraint. Note that precipitation values recorded for the period 1961-1990 at the nearest meteorological station (approximately 30km from Meerfelder Maar) show a mean annual precipitation of 908 mm/yr (Litt et al., 2009). We specified annual precipitation priors following a linear relationship between 500mm/yr at 11cal ky BP and 600mm/yr at 0 cal ky BP, with precipitation split amongst seasons following the modern seasonal distribution. The posterior January temperature and annual precipitation reconstruction obtained with this second test are shown in Figure 2.8.

For the period 0 to 3 cal ky BP, reconstructed temperature (Figure 2.8 top) is in good agreement with the first test (Figure 2.7 top) and prior and posterior precipitation are nearly the same. This implies that precipitation is not a constraining parameter for this period. For earlier periods January reconstructions for different experiments differ and prior and posterior precipitation for this second experiment differ too. This indicates that precipitation may have been somewhat higher than the values reconstructed by Litt et al. using the Bayesian Indicator Taxa method (Neumann et al., 2007) which is a Bayesian tuning of the Probability Density Function (pdf) method of Kühl et al. (2002).

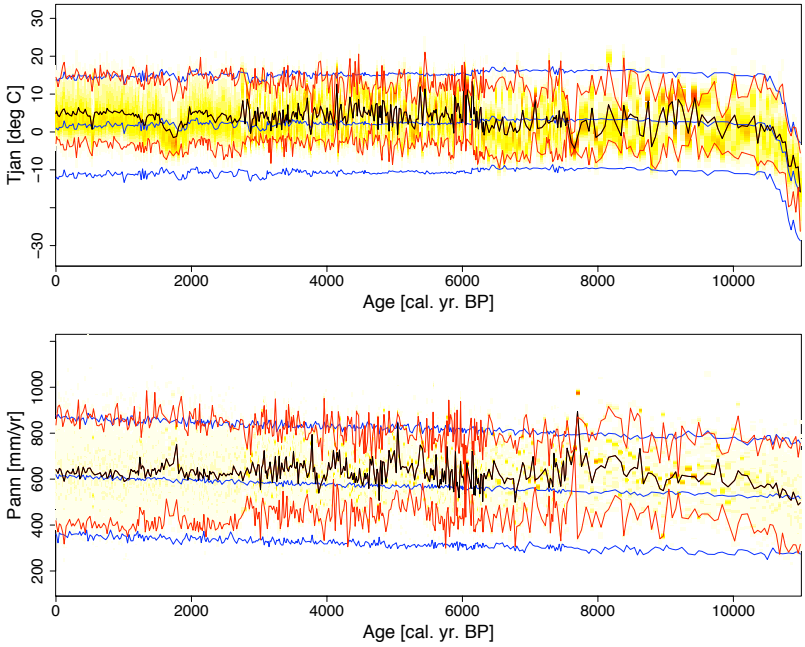

Figure 2.8: Climate reconstruction for Meerfelder Maar sequence during the Holocene using forced precipitation. Prior (blue lines): mean and 95% bilateral interval. Posterior (red lines) mean (highlighted with black) and 95% bilateral interval. The background colour (yellow to red) shows the posterior smoothed (low to high) density. For both plots, x-axis shows the age in cal yr BP, y-axis shows: (top) January reconstructed temperature in °C and (bottom) annual reconstructed precipitation in mm/yr.

2.4 Conclusion and discussion

Climate reconstruction by static inversion initiated by Guiot et al. (2000) is now used to take CO_2 variations into account (e.g. Wu et al., 2007a) and provides a method to reconstruct climate using different proxies (e.g. Hatté et al., 2009). The climate reconstruction by dynamic inversion retains these advantages and integrates a new generation of vegetation models. It is achieved using a Bayesian hierarchical model which sets the basis for causative modelling. We hope that this will encourage other work to use and extend this framework, and that the technical tools (statistical model and algorithm) presented in this article will help.

The use of a dynamic vegetation model has allowed an improvement of the "vegetation model inversion method" by including a temporal link. This allows a better exploitation of the information available from the fossil record. Further, the dynamic aspect of the model allows us to relax the assumption of equilibrium between vegetation and climate. In this paper, we focused on climate, but other variables are available as output of the vegetation model, e.g. primary production and carbon storage. These are useful for carbon cycle studies (see Wu et al., 2009) or could be used for paleo-fire modelling and more generally, studies of the past needing vegetation as input. The current study has, however, identified a number of problems listed below, and future work will concentrate on resolving these.

The validation tests showed that precipitation cannot be reconstructed as a four dimensional space. This seems to be partly due to the low sensitivity of the vegetation model to precipitation changes, and a sensitivity study is required. Until this problem has been addressed, we suggest that future use of this algorithm should use a single annual precipitation anomaly.

The results obtained using this method with the Meerfelder Maar sequence showed that the stability of the reconstructed values are linked to the pollen signal. When the pollen signal is nearly constant our reconstruction method over-amplifies small

variations in the pollen signal. When the pollen indicates two possible climates, the reconstruction may jump between two states, forming a non-continuous path of reconstructed values. However, when the changes occur in the pollen assemblages, the reconstructed variability is more coherent (low-frequency), for example during the period 3.7 to 0 cal ky BP. These differences are probably due to, (a) an over-fitting of the $p(Y|V)$ surfaces which force the vegetation model to follow the non-significant noise in pollen as if it were caused by climate change; (b) the absence of any direct time correlation in climate, which would result in a smoother reconstructed climate; (c) the use of a particle filter algorithm.

The $p(Y|V)$ model: We modelled the link between simulated vegetation and pollen using non-parametric surfaces. These surfaces are fitted using modern pollen and simulated vegetation. They poorly include the uncertainty we have on the link between vegetation and pollen. The inclusion of this uncertainty would summarise and transfer the incomplete knowledge of the link between vegetation and pollen, from the calibration to the reconstruction. This is a major departure from Bayesian modelling, and the next goal to improve the inversion method lies in using a parametric $p(Y|V)$ which allows the full propagation of uncertainty between calibration and reconstruction. The Bayesian framework for calibration and reconstruction in two separated steps has been presented in Haslett et al. (2006).

Temporal correlation in the climate: In opposition to the MAT reconstructions, which need to assume independence between samples, we model a vegetation link and obtain reconstructions that seem to be "noiser". As there is a dependency between samples, the vegetation model may require a larger climate change to fit both points than in the independent scheme. For the inversion of a dynamic model, it therefore seems essential to define a temporal climate correlation to counterbalance the effect of vegetation correlation, at least, when the resolution of the core is high. When the resolution of the core is low, correlation of the vegetation between samples is low and the reconstruction is nearly independent. However, the particle filter proposed here remains computationally more efficient and theoretically safer when using a dynamic vegetation model. Indeed, for a the static (or "at equilibrium") inversion of a dynamic

model, the experimenter has to run the model from nothing to the equilibrium (between vegetation and climate). This is called the burn-in phase and the diagnosis of convergence to equilibrium is always critical. With our method he can start from the vegetation reconstructed for the previous core point and has a fixed length of time to run the algorithm for: the time separating the two core points.

Filtering algorithm: The simple particle filter algorithm infers climate sequentially along the sedimentary sequence. However, it only optimises for coherence between the previous sample and the current one, and does not take into account the following sample. This feature can result in a chaotic path around the real climate. While some more complex algorithms attempt to minimise this problem, this remains an intrinsic problem of particle filtering, despite recent advances in the field (see for exemple `http://www-sigproc.eng.cam.ac.uk/smc/`).

The use of the filtering algorithm arises from the need to reduce the dimension of the climate space (equal to the number of samples times the number of climate variables). Since we cannot use another algorithm for the dynamic inversion, this is potentially an obstacle for: a. taking into account radiocarbon and other age errors and, b. the inference of any parameter that is dependent on the whole core. As the particle filter needs fixed stopping times (here the dates of the samples in the core) to simulate and weight particles, major changes and theoretical work would be required to integrate a possible error on this stopping time. Second, since the filter handles data sequentially, any parameter dependent on the whole core can only be estimated at the end of filtering. For example, this is a problem when considering a parameter θ of temporal correlation for the climate. At time t_2, the second point of the core, the algorithm has only the prior information to link t_1 and t_2 samples. The information about θ is updated along the core. This means that the reconstruction quality varies between points of the same core.

Despite the limits mentioned, the particle filter algorithm remains a promising tool for inversion of dynamic models or interpolation using such dynamic models. While work is required to further adapt this method, we believe that it will be a useful tool

in the field of climatology, in which a number of dynamic models are currently used or have been proposed.

Acknowledgements

The first author would like to thank Pascal Monestiez and Joel Chadœuf from BSP laboratory in INRA Avignon for fruitful discussions and statistical coaching. He was supported in computing work by Cyrille Blanpain and Philippe Dussouillez and by Judicaël Lebamba for the pollen diagram. The manuscript has been greatly improved thanks to the comments from three anonymous reviewers.

The authors are grateful for financial support for project DECVEG from the European Science Foundation (ESF) and Institut National des Sciences de l'Univers (INSU) under the EUROCORES Programme EuroCLIMATE. This work is also a contribution to the project PICC funded by the French Agence Nationale de la Recherche (programme blanc).

Chapter 3

A Multinomial Poisson model for spatial data with structural zeros: *European-scale linkage of simulated vegetation and pollen data for palaeoclimatology*

This chapter is submitted to *Environmetrics* as the following research paper

Garreta, V., Guiot, J., Mortier, F., Chadœuf, J., and Hély, C., (submitted) A Multinomial Poisson model for spatial data with structural zeros: *European scale linkage of simulated vegetation and pollen data for the palaeoclimatology* submitted to *Environmetrics*

I realised the work presented in this chapter with the help of Frédéric Mortier (CIRAD, Montpellier) and Joël Chadœuf (INRA, Avignon) for the mathematical development of the Multinomial-Poisson model. I wrote the chapter.

Abstract Palaeoclimate reconstructions are usually based on statistical relationships linking sediment pollen assemblages and climate. These statistical models are calibrated using a modern, typically massive and spatial, dataset. Replacing these statistical models by mechanistic vegetation models improves past climate prediction but needs the coupling of the vegetation model outputs and pollen data. We propose here a (statistical) process-based model for such a coupling, which takes into account the pollen spatial dependencies.

We consider two major difficulties. First, the multinomial pollen data present an overdispersion and structural zeros, which require modelling. Second, the model has to allow its inference on a massive dataset since it must fit for a large climate range.

In the hierarchical model framework, we develop a Bayesian model based on the four main processes: pollen production, spatial dispersion, accumulation and sampling. Accumulation and sampling processes are modelled using a Multinomial-Poisson model that allows for overdispersion and structural zeros (null multinomial probabilities). The dispersion is modelled using a Gaussian kernel, and the vegetation model errors are described using a mixture model.

We demonstrate that MP parameters are identifiable and apply this result for the inference of several simulated datasets. We finally perform inference on the European pollen dataset, which is made possible by the parallelisation of the Monte Carlo Markov Chain algorithm.

We build three diagnostics to investigate the adequacy of the hierarchical model at its various levels. These diagnostics detect that the MP model is not sufficient to represent all the multinomial overdispersion present in the European dataset. We therefore discuss in the conclusion several ways to model such overdispersion and our preferred one is to over-disperse the Poisson distribution using a Negative Binomial distribution.

3.1 Introduction

Pollen-based palaeoclimate reconstructions are obtained using a statistical model of the relationship between pollen assemblages sampled in sediment cores and climate (e.g. the pioneering work of Webb and Bryson (1972) and Brewer et al. (2007); Guiot and De Vernal (2007) for a recent review of the methods). These statistical models are called Transfer Function (TF). The reconstruction process is twofold. The *calibration* step consists in fitting the TF to a modern dataset that is typically massive (over 1000 samples) and spatially distributed (over Europe in our case). This is imposed by the need to infer a robust link between climate and pollen over a large climate range, at least as large as expected for past variations. The *reconstruction* step consists in the *inversion* or *prediction* of the climate based on ancient pollen, depending on whether the TF is direct - pollen = f(climate, ϵ) - or backward - climate = f(pollen, ϵ) - with ϵ an error term.

Guiot et al. (2000) introduced the idea of embedding a mechanistic vegetation model into the direct form of the TF. They called their prediction method 'inversion of the vegetation model' because it imposes the inversion of the model's computer code for climate reconstruction. This class of TF provides a new picture of past climate and vegetation compared to the purely statistical TF by integrating the higher level of complexity of vegetation modelling. It also includes additional environmental variables such as CO_2 atmospheric concentration, which may result in large deviations compared to classical TF reconstructions (Wu et al., 2007a). But until now, its - potentially - major advantage remains not exploited. Indeed, this new class of TF offers the opportunity to develop a full process-based modelling of the link between climate and pollen. Such modelling, in addition to being causative, better understandable and controllable, could resolve the 'no-analogue' problem that is the lack of reconstruction power of the classic TF outside the modern climate range. It is an extrapolation problem inherent to TF that are not process-based and thus, not reliable outside the (modern) climate range used for calibration.

Our objective is to develop a new TF linking the vegetation model LPJ-GUESS

(Smith et al., 2001) and the modern pollen dataset over Europe. Its building raises two major challenges. First, we have to consider the modelling of multinomial overdispersion in the context of many zeros. Indeed, the pollen counts that are naturally modelled using the multinomial distribution (Haslett et al., 2006) are over-dispersed (i.e. with a variance larger than the one of the multinomial distribution) and contain 'structural' zeros in the sense that the probabilities controlling the multinomial, for a certain component, can be exactly zero. These zeros are due to the true absence of the taxa in the sampled region or result from a process (e.g. accumulation) generating them. The second challenge consists in building a model with a sufficient complexity to represent the processes linking vegetation and pollen but whose inference remains feasible.

We propose a Multinomial-Poisson (MP) model to represent the multinomial overdispersion. In this model, we define the probabilities of the multinomial distribution as normalised sampling from k Poisson variables, one per multinomial component. This original model is a discrete version (for the variable generating the multinomial probabilities) of the well-known Multinomial-Dirichlet model (MD, e.g. Leonard, 1977). The use of the discrete Poisson distribution allows structural zeros in the sense that multinomial probabilities can be null. It has the same number of parameters as the MD model (one per component of the composition). We demonstrate that $k-1$ parameters are related to the expectation of the multinomial probabilities and the kth controls the variance of the proportions, i.e. the overdispersion of the multinomial.

To answer the first challenge, we propose a hierarchical model whose levels are the main (natural) processes linking the vegetation model outputs and pollen data. These processes include the anthropogenic disturbance of the vegetation simulated by LPJ-GUESS, pollen production, dispersion, accumulation and sampling. The MP model presented above emerges from the modelling of the two later processes (sampling given accumulated pollen). The spatial feature is induced by the modelling of pollen dispersal.

The dimension of the proposed hierarchical model is high. If we note k the number of taxa (the components of the multinomial, $k = 15$ for our European dataset) and n, the number of sites ($n = 1301$ for Europe), our model contains $5 * k$ parameters and

$2 * k$ latent fields sampled at n sites. We work under the Bayesian paradigm and use a Monte Carlo Markov Chain (MCMC, e.g. Robert and Casella, 1999) algorithm for inference. The inference is made possible thanks to our model's structure that allows to parallelise at each step the Metropolis-within-Gibbs algorithm we implemented.

A necessary step of this applied work is to check the model adequacy, i.e. to investigate the *a posteriori* consistency between the model structure and (hidden) data structure. The hierarchical Bayesian model is defined by three levels and its inference is so computationally demanding that it is performed once and for all. Therefore, we have to use methods based on the posterior simulations such as the posterior predictive p-value (Gelman et al., 1996). Since these methods are known to be imperfect (i.e. conservative) we combine three of them to investigate the adequacy of each level.

In section 3.2 we present the different levels of the hierarchical model and we consider the identifiability of the Multinomial-Poisson model parameters (demonstration in the appendix B). In section 3.3 we explain the parallelisation of the MCMC algorithm and develop three diagnostics for the model adequacy testing. Inference tests with simulated datasets are presented in section 3.4.1 and model adequacy is tested. The complete European dataset is inferred in section 3.4.2 and model's adequacy is discussed. We end the paper by a discussion around the Multinomial-Poisson modelling and perspectives for this model in palaeoclimatology.

3.2 Process-based model

We build a hierarchical Bayesian model representing the main processes linking LPJ-GUESS simulated vegetation and pollen data sampled over Europe. Each of these processes is modelled as a hidden level and their succession is causative: starting from the simulated vegetation → actual vegetation → produced and dispersed pollen → accumulated pollen → sampled pollen. In the description of the levels' distribution we will explicitly note the underlying hypothesis. The later two processes (accumulation and dispersal) are naturally modelled following a Multinomial-Poisson model whose

moments of the proportions p are calculated in the appendix B.

We note $Y_i = (Y_i^1, .., Y_i^k)$ a multinomial vector of the pollen assemblage sampled at site $i = 1..n$. Y_i^j is the number of pollen grain of the taxa j at site i. The vegetation data, simulated at the same sites that pollen samples, are noted as NPP_i=(NPP_i^1,..,NPP_i^k). NPP_i^j is the (absolute) net primary production simulated by LPJ-GUESS at site i for the taxa j.

3.2.1 From potential to actual vegetation: mixture model

The vegetation simulated using LPJ-GUESS is *potential* vegetation. This means that it is controlled by climate, soil properties, CO_2 and not disturbed by human activities. Thus, the modern vegetation composition is expected to be a noisy image of the simulated NPP. We model this modern vegetation that products pollen, $V_i = (V_i^1, .., V_i^k)$, termed 'actual' vegetation, as a hidden variable with conditional distribution:

$$\left[V_i^j | \text{NPP}_i^j, \sigma^j, m^j, q^j\right] = \begin{cases} \left[V_i^j | \text{NPP}_i^j > 0\right] = \mathcal{G}\left(\frac{(\text{NPP}_i^j)^2}{\sigma^j}, \frac{\text{NPP}_i^j}{\sigma^j}\right) \\ \left[V_i^j | \text{NPP}_i^j = 0\right] = q^j \delta_0 + (1 - q^j)\mathcal{G}\left(\frac{(m^j)^2}{\sigma^j}, \frac{m^j}{\sigma^j}\right) \end{cases}$$

where δ_0 is the Dirac mass at 0 and $\mathcal{G}(s, r)$ the gamma distribution with shape and rate parameters s and r. This modelling of the anthropogenic disturbance is interpreted as follows: when the taxa j is simulated at site i ($\text{NPP}_i^j > 0$), the vegetation is distributed following a gamma distribution (showing no probability mass at 0) centred on the simulated value NPP_i^j with variance σ^j. In doing so, we assume that, if potentially present, a species cannot be eradicated totally by mankind. It is, at least, present in a very small proportion around the considered site i. When not potentially present at site i ($\text{NPP}_i^j = 0$), the actual vegetation taxa V_i^j has a probability q^j ($\in [0, 1]$) of being absent. If present regardless of $\text{NPP}_i^j = 0$ (e.g. because mankind planted it) it is distributed following a gamma distribution centred on m^j and with variance σ^j. The overall mean and variance of $\left[V_i^j | \text{NPP}_i^j = 0\right]$ are $(1 - q^j)m^j$ and $(1 - q^j)(\sigma^j + q^j m^j)$.

This representation of the anthropogenic pressure on the vegetation over Europe is based on the hypothesis of stationarity over the spatial domain. In studies involving

spatial descriptors of the anthropogenic disturbance such descriptors could be used as regressors for the q, m and σ parameters.

3.2.2 Linear production and Gaussian dispersion of the pollen

Each species j produces an absolute quantity of pollen linearly related to its abundance, here expressed as NPP. This assumption is commonly accepted in palaeo-ecology (e.g. Sugita, 2007a) and basically lies on the following approximation: twice more trees produce twice more pollen. We model the absolute quantity of pollen produced at site i by species j

$$b^j.V_i^j$$

The pollen produced by each species is dispersed following a Gaussian kernel whose dispersal length parameter depends on the species. For each spatial location i, the pollen brought by dispersal is

$$b^j S_i^j = b^j \sum_{l=1}^{n} \alpha^j \left(d(i,l)\right) V_l^j$$

equal to the convolution of the Gaussian kernel $\alpha^j(.)$ centred on i. The kernel is $\alpha^j(x) \propto \frac{1}{\gamma^j} \exp\left(-\frac{x^2}{2(\gamma^j)^2}\right)$. The distance $d(x,y) = \sqrt{(x-y)^2}$ is Euclidian and γ^j $(j = 1..k)$ are the dispersal distance parameters.

This representation of the production and dispersion processes is based on the hypothesis of stationarity of these processes over the spatial domain. In other studies, the production parameters b and kernels $\alpha(.)$ could vary spatially, i.e. become some $b(s)$ random fields and $k(.,s)$ kernels as, for example, in Higdon (1998).

3.2.3 Accumulation and sampling of the pollen: Poisson-Multinomial model

The pollen accumulation in natural traps (mosses, lakes, peat bogs) results from many different processes (anisotropy of the local dispersal function, heterogeneity of the trap

capture, differential concentration of the incoming fluxes, etc). Information about the trap is often lacking or loose (e.g. the size of the lake in the past, the strength of past wind fields). We chose to model accumulation process globally, as a random Poisson distribution, instead of modelling its too numerous and complex underlying processes. The pollen X_i^j of the species j accumulated at the site i is

$$[X_i^j | b^j, S_i^j] = \mathcal{P}(b^j S_i^j)$$

where $\mathcal{P}$ is the Poisson distribution. Note that when the pollen is theoretically absent ($S_i^j = 0$) the Poisson distribution degenerates into a Dirac mass at 0. Thus, no pollen can appear during the accumulation process if it was not brought by dispersion.

The sediment accumulated in the trap is sampled; pollen taxa are recognised and counted. Typically, palynologists count a number N_i of pollen grains depending on various criteria such as sample quality and diversity. We model the sampling process (including also recognising and counting) following the multinomial distribution

$$[Y_i | X_i, N_i] = \mathcal{M}(p_i, N_i)$$

where $p_i = (p_i^1, .., p_i^k)$ with $p_i^j = X_i^j / \Sigma_{m=1..k} X_i^m$ is the *in situ* proportion of pollen j accumulated in the trap located at i. The use of the multinomial distribution is classical in palynology and neglects the error coming from recognising and counting (used e.g. in Prentice and Parsons (1983) it has been recently discussed in Haslett et al. (2006)).

This way of modelling palynological overdispersion with respect to the multinomial distribution and in the presence of structural zeros (null p_i^j) is original and parsimonious. This model is based on the structure of the Multinomial-Dirichlet (MD) model and uses a discrete Poisson distribution to account for zeros in the proportions p^j (due to $X^j = 0$). The MD model is usually generated as follows:

- simulate k independent $X^j \sim \mathcal{G}(\text{shape} = \beta^j, \text{rate} = 1)$
- compute the proportions p with $p_j = X^j / \Sigma_{m=1..k} X^m$

- given p and N, generate Y following $[Y|p, N] = \mathcal{M}(p, N)$

In the MD model, the probability for $p^j = 0$ is null since the gamma distribution is continuous. Bayesians use the conjugacy between the Dirichlet distribution (defined by the ratio of the gamma random variables) and the multinomial distribution to sample from the posterior of the β^j parameters. Results about the moments of the Dirichlet distribution prove that the β^j parameters are identifiable and can be interpreted in terms of mean and variance of the Dirichlet distribution. For our model, involving ratio of Poisson distributions, we calculate the mean and variance of the ratios. These results show that the model is identifiable and allow the interpretation of the parameters b^j. Main results are presented in the following section and the complete demonstration is given in the appendix B.

Identifiability of the Multinomial-Poisson model

The identifiability of the k b^j parameters is not trivial due to the normalisation. For any site i the pollen accumulated, X_i, is distributed following Poisson distributions centred on the k-dimensional mean vector $(b^1 S_i^1, .., b^k S_i^k)$. S_i^j are the spatial regressors and b^j some species specific parameters. The normalisation of the X_i^j forms the k proportions p_i^j. The k proportions define a $k-1$ dimensional space due to their sum to one constraint and they are matched to the pollen data throughout the multinomial likelihood. Therefore, *a priori*, the $k-1$ dimensional space for the proportions cannot constrain the k b^j parameters. This is demonstrated by calculating the p^j expectation (see appendix B)

$$\mathbb{E}[p_i^j | \Sigma_{j=1..k} X_i^j > 0] = \frac{S_i^j . b^j}{\Sigma_{j=1..k} b^j S_i^j}$$

which is centred on the same value for any vector $\xi * (b^1, .., b^k)$ with a real $\xi > 0$.

We show that the kth parameter is related to the variance of the proportions. Let us leave the site subscript i and note $a^j = S^j . b^j / \Sigma_{j=1..k} b^j S^j$ and $K = \Sigma_{j=1..k} b^j S^j$. Thus the a^j's are $k-1$ parameters (due to the sum to one constraint) that are indentifiable by

the mean of the proportion data and K is the overall sum parameter. We demonstrate in the appendix B that for a not too little K, say $K \geq 10$ then

$$\mathbb{V}\text{ar}\left[p^j | \Sigma_{j=1..k} X^j > 0\right] \approx \frac{a^j(1-a^j)}{K}$$

This result implies that the parameter K is identifiable by the variance of the proportion data. It is the parameter, in the Multinomial-Poisson model, which accounts for overdispersion. When this parameter is very large the overdispersion disappears. In the Bayesian framework for inference, one can restrict the prior for K to a maximum value corresponding to a numerically insignificant overdispersion to allow convergence of the MCMC chain in the absence of overdispersion.

3.2.4 Priors

We list here the priors selected to complete the Bayesian model.

• The parameters of pollen production per species (b^j) are independent gamma distributions.

$$[b^j] = \mathcal{G}(10^{-3}, 10^{-3})$$

Because no prior arises intuitively from the problem at hand, we chose the gamma distribution, which is conjugated to the Poisson distribution and allows use a step of Gibbs sampling for the MCMC algorithm. The gamma parameters are selected to form a weakly informative prior.

• The parameters γ^j of dispersal distance are difficult to estimate. Indeed, no conjugacy property exist and for every new proposed value γ^{j*} it is necessary to compute the (pollen dispersal) kernel for every site, which is prohibitive with respect to computing time and memory size. Based on Diggle et al. (2003) in geostatistics, we propose to pre-compute L different kernels associated to L values $g_1, .., g_L$ of the dispersal distance. The inference algorithm consists in scanning this discretised space. The prior

underlying this method is a discrete uniform prior over the $g_1, .., g_L$ pre-specified values.

$$[\gamma^j] = \Sigma_{l=1}^{L} \delta_{\gamma^j = g_l} / L$$

These values are chosen so that the grid fully covers the prior range of γ. Moreover the distance between consecutive g_l values must not be too large since it results in poor approximation of the posterior density due to its coarse discretisation and poor mixing of the MCMC chain. Indeed, large distances between consecutive g_l make very different the consecutive kernels, which lead to high rejection during the Metropolis step. In practice we use a grid covering the range of possibilities for γ with a uniform grid lag determined by the number of kernel matrices that can be computed using a reasonable amount of memory size.

• The mixture parameters q^j for the actual vegetation are assumed to be independent and uniformly distributed over $[0.5, 1]$.

$$[q^j] = \mathcal{U}(0.5, 1)$$

The lower bound for this distribution is based on the assumption that, over the large area considered (Europe), using a vegetation model allows to better predict the absence of the vegetation than a pure random experiment which has a probability of 0.5 to be right.

• Finally, scale and shape parameters m^j and σ^j for each species are independent and follow Gaussian distributions trucated to be strictly positive.

$$[m^j] = \mathcal{N}(h_1^j, h_1^j/2) \qquad \text{truncated to} \qquad (0; +\infty)$$

The parameter m^j is the mean of the actual vegetation taxa j when it is present but has not been simulated by the model. The vegetation absolute quantity in this situation (presence regardless of the potential absence) is hard to assess because this means that mankind facilitates or directly helps the establishment and growth of the taxa. h_1^j is set equal to the mean of the simulated NPP^j over Europe when present ($\text{NPP}^j > 0$).

$$[\sigma^j] = \mathcal{N}(h_2^j, h_2^j/2) \quad \text{truncated to} \quad (0; +\infty)$$

h_2^j is set equal to the variance of the simulated NPP^j over Europe when present ($\text{NPP}^j > 0$).

3.3 Inference and model checking

3.3.1 Inference method using computer parallelism

The hierarchical Bayesian model described in the preceding section contains two sets of $k = 15$ latent fields (V^j and X^j with $j = 1..k$) and five sets of k latent variables (q^j, m^j, σ^j, b^j and γ^j with $j = 1..k$) to be inferred. We use a Metropolis-within-Gibbs algorithm (e.g. Robert and Casella, 1999) to sample from the posterior of each single parameter and each point of the latent fields (1301 points by field). This means that we sample, in turn, new values of the parameters and points following their full conditional distributions. Computing time is critical because the number of variables inferred is large and the spatial feature of the model requires computation of many matrix-vector products. More than 99% of computing time is devoted to sampling from the full-conditional distribution of points from the V^j fields. Indeed the full conditional distribution of a V_i^j point is

$$[V_i^j|...] \propto \left(\prod_{k=1}^{N}[X_k^j|b^j S_k^j]\right)[V_i^j|\text{NPP}_i^j, q^j, m^j, \sigma^j]$$

which is time demanding to compute since S_k^j is a convolution of the whole field V^j. By remarking that the full-conditional distributions of the field j's points only depend on the field j (and related variables), we can parallelise the algorithm over the fields, i.e. simulate independently each field. In addition to the parallelism, the variables V_i^j can be sampled in blocks $V_{k_1}^j, ..., V_{k_n}^j$ to reduce the number of convolutions needed to sample all the field j. But, contrary to parallelism, this speed increase is counterbalanced by a potentially slower convergence of the algorithm: the rejection in the Metropolis step increases with the blocks' size.

The inference algorithm is coded using C language. Parallelisation of the C code on one single (multi-core) shared-memory computer is obtained adding a few lines of OpenMP language (`http://openmp.org`). The only difficulty when parallelising the code is the need for a parallel and efficient random number generator. We use the combined multiple recursive random number generators from L'Ecuyer (1999) whose 'RNGstreams' C code is freely available (`http://www.iro.umontreal.ca/~lecuyer/`).

The reported computation times are obtained using 7 processors of a 64bit computer composed of two quad-cores at 2GHz.

3.3.2 Bayesian checking of a huge hierarchical model

We want to check the consistency between model structure and (hidden) data structure, often referred to as 'goodness of fit' or 'model adequacy' testing. More specifically, such tests are expected to detect cases where there is more dispersion in the data than specified by the model and incoherence between model and data's distributions. Since the model is hierarchical (with two hidden levels) and contains a non-Gaussian structure, there is no well-defined way to test or at least measure its adequacy. We define a Bayesian model 'checking' (in the spirit of Gelman et al., 1996; Stern and Cressie, 2000; Marshall and Spiegelhalter, 2003) by a series of diagnostics to check whether or not the model's parametrical structure fit the data in its various levels.

For that purpose, three diagnostics are presented to check the different levels of the hierarchical model. These diagnostics are based on two main ideas: first, the comparison between *discrepancies* (test statistics depending on the model's parameters, Gelman et al., 1996) obtained from the posterior simulations versus those obtained from reference simulations defined by the model structure. Second, since the whole hierarchical model inference cannot be re-run for a cross-validation exercise we use the concept of mixed replications (Marshall and Spiegelhalter, 2003, 2007), which theoretically improves the power of the diagnostics in detecting inadequacies. The approach and the three different measures of discrepancies are presented in the following sections.

Likelihood level check: Posterior predictive p-value

The idea behind this method comes from Guttman (1967). It has been formalised by Rubin (1984) and is deeply discussed in Gelman et al. (1996).

At the end of the inference we have a set of M posterior simulations ($X^{\text{post},m}$, $V^{\text{post},m}$, $b^{\text{post,m}}$, $\gamma^{\text{post},m}$, $\sigma^{\text{post},m}$, $m^{\text{post},m}$, $q^{\text{post},m}$) with $m = 1..M$ following

$$[X, V, b, \gamma, \sigma, m, q|Y] \propto [Y|X]\,[X|b, \gamma, V]\,[V|\sigma, m, q]$$

From the posterior simulations $X^{\text{post},m}$ ($m = 1..M$) a set of corresponding 'replicate' data Y^{rep} can be simulated following the multinomial model

$$Y^{\text{rep},m} \sim [Y|X^{\text{post},m}]$$

The posterior predictive p-value diagnostic consists in selecting one discrepancy measure $T(Y, X)$ and comparing the distributions of $T(Y, X^{\text{post},m})$ and $T(Y^{\text{rep},m}, X^{\text{post},m})$ through the p-value

$$p(T(Y, X^{\text{post},m}) < T(Y^{\text{rep},m}, X^{\text{post},m})) \tag{3.1}$$

which is approximated as the proportion of times that $T(Y, X^{\text{post},m})$ is lower than $T(Y^{\text{rep},m}, X^{\text{post},m})$ for a set of M posterior simulations $X^{\text{post},m}$ ($m = 1..M$). If the p-value is lower or higher than pre-specified bounds, e.g. 0.025 and 0.975, the posterior discrepancy is said to be outside of its reference distribution defined by the model through the replicates. This indicates that some traits of the posited model computed over unconstrained simulations (the replicates) are significantly different from the same traits computed over the posterior simulations.

We use T_1, the deviance of the multinomial distribution, as the discrepancy measure

$$T_1(Y, X) = -2 \sum_{i=1}^{N} \log(\mathcal{M}(Y_i, p_i, N_i)) \tag{3.2}$$

The posterior predictive p-value is conservative, i.e. it fails in detecting small to medium inconsistencies between model and data (e.g. Stern and Cressie, 2000). This

comes from the fact that X_i^{post} is influenced by Y_i (through the likelihood) and thus Y_i does not appear inconsistent regarding $[Y_i|X_i^{\text{post}}]$ distribution. Moreover this diagnostic only checks for inconsistencies measurable at the likelihood level.

Likelihood and first level check: Mixed posterior predictive p-value

The idea of mixed predictive distribution from Marshall and Spiegelhalter (2003) is the following: instead of using the reference distribution $Y^{\text{rep}} \sim [Y|X^{\text{post}}]$, one can simulate and use the reference distribution for a higher level, here, $(Y^{\text{rep},m}, X^{\text{rep},m}) \sim [Y|X^{\text{rep},m}]\,[X^{\text{rep},m}|b^{\text{post}}, S^{\text{post},m}]$, which is less constrained by Y (only through S^{post}). One can then compute the quantiles of each Y_i for the reference distribution defined by Y_i^{rep}, the marginal distribution of $(Y_i^{\text{rep}}, X_i^{\text{rep}})$. Since there is no consensus on the definition of multivariate quantiles, we propose to extend the idea of posterior predictive distribution and compute the p-value

$$p(T(Y, X^{\text{post}}, b^{\text{post}}, S^{\text{post}}) < T(Y^{\text{rep}}, X^{\text{rep}}, b^{\text{post}}, S^{\text{post}})) \tag{3.3}$$

approximated as the proportion of times that discrepancy $T(Y, X^{\text{post},m}, b^{\text{post}}, S^{\text{post},m})$ is lower than $T(Y^{\text{rep},m}, X^{\text{rep}}, b^{\text{post}}, S^{\text{post},m})$ for $m = 1..M$ simulations.

We use the discrepancy T_2, sum of the deviances of multinomial and Poisson distributions

$$T_2(Y, X, b, S) = -2\sum_{i=1}^{N}\left(\log(\mathcal{M}(Y_i, p_i, N_i)) + \sum_{j=1}^{k}\log(\mathcal{P}(X_i^j, b^j S_i^j))\right)$$

Using this discrepancy we compare traits of the reference and posterior distributions for both likelihood and first hidden level.

Full model check: Full mixed posterior predictive p-value

We extend the preceding idea to the whole hierarchical model by generating replicate distributions for both hidden levels and data. From the posterior simulation of the

parameters $\theta^{\text{post},m} = (b^{\text{post},m},\ \gamma^{\text{post},m},\ \sigma^{\text{post},m},\ m^{\text{post},m},\ q^{\text{post},m})$ we simulate a replicate $(Y^{\text{rep},m},\ X^{\text{rep},m},\ V^{\text{rep},m})$ following

$$(Y^{\text{rep},m}, X^{\text{rep},m}, V^{\text{rep},m}) \sim [Y|X^{\text{rep},m}]\,[X^{\text{rep},m}|b^{\text{post},m}, S^{\text{rep},m}] \\ [V^{\text{rep},m}|\text{NPP}, \sigma^{\text{post},m}, m^{\text{post},m}, q^{\text{post},m}]$$

Using the replicates and posterior simulations we compute the p-value

$$p(T(Y, X^{\text{post}}, V^{\text{post}}, \theta^{\text{post}}) < T(Y^{\text{rep}}, X^{\text{rep}}, V^{\text{rep}}, \theta^{\text{post}})) \tag{3.4}$$

with the discrepancy T_3, which is the deviance of the entire model

$$\begin{aligned} T_3(Y, X, V, b, \gamma, \sigma, m, q) = &-2\sum_{i=1}^{N} \log(\mathcal{M}(Y_i, p_i, N_i)) + \\ &- 2\sum_{i=1}^{N}\sum_{j=1}^{k} \left(\log(\mathcal{P}(X_i^j, b^j S_i^j)) + \log([V_i^j|\text{NPP}_i^j, \sigma^j, m^j, q^j])\right) \end{aligned} \tag{3.5}$$

3.4 Application to simulated and European dataset

3.4.1 Simulated datasets

We made several inference tests using simulated datasets to check the correctness of the computer code, the robustness of the MCMC algorithm and the influence of vegetation priors. The datasets are simulated following the model for three ($j = 1..3$) different species sampled at 150 points ($i = 1..150$) distributed following a uniform distribution on a one-dimensional space (between 0 and 40), see Figure 3.1.

For a selected set of parameter values $(\hat{r}, \hat{b}, \hat{\gamma}, \hat{q}, \hat{m}, \hat{\sigma})$ a 'toy' dataset is simulated as follows:

- NPP_i^j are simulated following a spatially structured Gaussian field truncated at 0 (simulations below 0 are set to 0). The spatial structure is given by a Gaussian covariance with a scale parameter $\hat{r}^j$ by species and a variance equal to 1.

- Actual vegetation V_i^j is simulated following the mixture model with selected $(\hat{q}, \hat{m}, \hat{\sigma})$ parameter values.
- The pollen dispersed at each point for each species, S_i^j, is computed with selected $\hat{\gamma}$ parameter values.
- Accumulated pollen X_i^j is simulated following the Poisson distribution with selected $\hat{b}$ parameter values.
- After normalisation, pollen accumulated proportions p_i^j are computed and 'sampled' pollen is generated following a multinomial distribution whose total counts (N_i) is equal to 200.

As in the real world, inference is performed using only sampled pollen and simulated NPP.

We made two kinds of inference tests.

First, for different sets of parameters ($\hat{r}^j \in \{1; 2; 3\}$, $\hat{b}^j \in \{10; 50; 100\}$, $\hat{\gamma}^j \in \{1; 2; 3\}$, $\hat{q}^j \in \{0.6; 0.7; 0.8; 0.9\}$, $\hat{m}^j \in \{0.5, 1, 2\} * \text{mean}(\text{NPP}^j[\text{NPP}^j > 0])$ and $\hat{\sigma}^j = \text{var}(\text{NPP}^j[\text{NPP}^j > 0])$) we ran the inference algorithm (for 1.5 million MCMC iterations in 5h) and checked that the 95% Highest Posterior Regions (HPR) contained the parameters and V, X latent fields used for the simulation in approximately 95% of the cases. These tests indicate that the algorithm is robust even with large overdispersion from the Poisson latent field. We present here the most overdispersed case ($b = 10$) Figure 3.1).

The second kind of test (not presented here) consisted in using wrong informative priors for the vegetation parameters m and σ to check their effect on the inferred values. This test can be interpreted as a rough prior sensitivity analysis to asses if the informative prior, used for the real dataset, will have a strong influence on the inferred values of all the parameters. The 'wrong' priors used for these tests are the ones described section 3.2.4 with both of their parameters multiplied by $\{0.3; 0.5; 2; 3\}$. Results indicate that these priors only influence the inferred m, σ and q values. Moreover, when

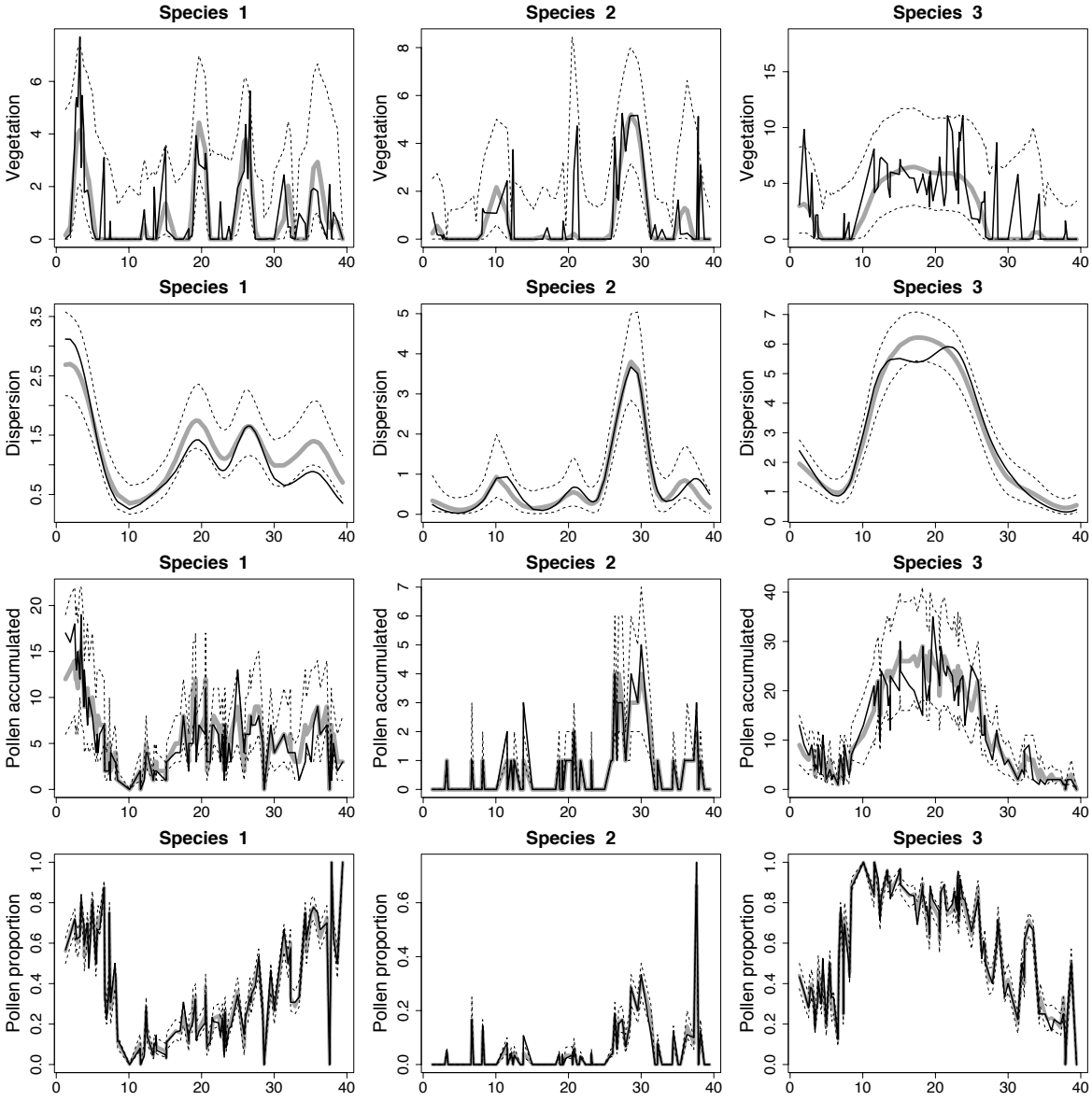

Figure 3.1: Inference for a dataset simulated over an arbitrary spatial index (x-axis). Each column is a different species and each row a different latent quantity. The black line is the expected, simulated value. Posterior mean is the thick line in light grey and 95% Highest Posterior Region (HPR) are given by the dashed lines. First line shows the absolute vegetation abundance per species V^j ($j = 1..3$). Second line shows S^j, the dispersed vegetation or dispersed pollen without reference to the relative production carried by b. Third line shows the accumulated pollen X^j and last line the multinomial probabilities p^j for each sampled site.

the prior range does not include the expected value the posterior mean is as close as possible of it, but stays in 95% high probability region of the prior.

3.4.2 European dataset

Data

The modern pollen database compiled by Bordon (2008) consists of pollen grain abundances recognised per taxa. Their total (N_i) usually ranges between 100 and 500. A vegetation model output is the Net Primary Production (NPP in $kg.m^{-1}.yr^{-1}$) per species considered in the model. We use 17 major tree species in Europe plus one group representing all grasses and shrubs. To be in agreement, model species and pollen taxa are reduced to 14 tree taxa and a grass/shurb group (see Table 3.1). For a full description in term of vegetation model parameters, see Miller et al. (2008) and Garreta et al. (2009). To simulate the vegetation corresponding to modern pollen samples we run the vegetation model for each pollen sample site using the 20th century climate dataset CRU TS1.2 (New et al., 2002), available at the monthly time-step and spatial grid of 10' resolution. These series have been interpolated at the pollen sites using ordinary kriging with the altitude as external drift (e.g. Cressie, 1991).

Then, for each site i,

1. we spin-up the model during 500 years using a climate chronology which is detrended and the CO_2 concentration of 1901,
2. we run the model for the years 1901-1990 using the interpolated climate times series and CO_2 measured for this period,
3. we retain the average NPP for the years 1961-1990 to form NPP_i.

j	Pollen group	Vegetation group based on model outputs
1	`Abies`	`Abi_alb`
2	`Alnus`	`Aln_inc`
3	`Betula`	`Bet_pen` + `Bet_pub`
4	`Carpinus`	`Car_bet`
5	`Corylus`	`Cor_ave`
6	`Fagus`	`Fag_syl`
7	`Fraxinus`	`Fra_exc`
8	`Picea`	`Pic_abi`
9	`Pinus`	`Pin_syl` + `Pin_hal`
10	`QuercusE`	`Que_coc` + `Que_ile`
11	`QuercusD`	`Que_rob`
12	`Tilia`	`Til_cor`
13	`Ulmus`	`Ulm_gla`
14	`Populus`	`Pop_tre`
15	`GrSh`	`C3_gr`

Table 3.1: The fifteen groups ($j = 1..15$) of pollen and vegetation model outputs.

Results

The inference algorithm is run by successive sequences of 50k (50,000) iterations (20 h long). This allows to start the treatment of the MCMC chain values before the end of many days of computing, to check for coding and other errors and to monitor convergence. For convergence monitoring, classical tests such as the comparison between/within variance of multiple MCMC chains (Gelman and Rubin, 1992) is not available since we work with only one MCMC chain. We check visually the stationarity of each parameter's chain (see Figure 3.4 for a subset of such outputs). We use the deviance of the multinomial as an indicator of lack of convergence. Indeed, for the first 400k iterations, each 50k-long sequence showed a significant decrease in the deviance

criteria computed over the posterior simulations (Eq. 3.2). We use one iteration over 100 between iterations 500k and 600k to compute the following model checks, output summaries and plots.

Model Checking

The posterior predictive p-value for the deviance, presented in Equation 3.1, is equal to 1. This indicates that the distribution of the posterior simulations' deviance is significantly higher than its reference distribution, e.g. its mean is 22% higher than the one of its reference. This indicates that the model is not adequate, at least, at its likelihood level.

The mixed posterior predictive p-value (Eq. 3.3), which measures likelihood and first (X) level coherence with data, is equal to 1. Thus, the likelihood and first level are incoherent with data, e.g. the mean of the posterior deviance is 24% higher than the one of its reference.

Finally, the full mixed posterior predictive p-value (Eq. 3.4) is equal to 0.96 and thus, close to being considered incoherent (for our 5% two-sided test). We decompose this distribution following the three terms of Equation 3.5 and plot them on Figure 3.2. The graph clearly indicates that likelihood level is not coherent nor is the first hidden level and this is compensated by the vegetation level that has a low deviance score.

We found two different explanations for the lack of fit of the model. One problem is the lack of overdispersion modelled by the likelihood and first hierarchical level. This is indicated by the discrepancy values that pointed a *strong* model-data incoherence at these levels, combined with a high inferred overdispersion for the MP model. To illustrate this point we selected a point i_1 whose mean discrepancy is one of the highest compared to the reference. The values of data and posterior hierarchical levels for a selected subset of taxa of this point are shown in Table 3.2.

Table 3.2 does not show indisputably that posterior proportions means are more

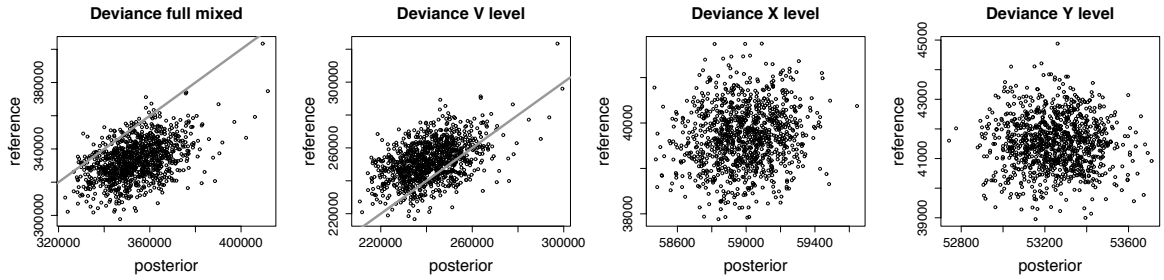

Figure 3.2: Decomposition of the the full mixed posterior deviance. Each graph presents the deviance obtained from the posterior (x-axis) versus the reference deviance distribution (y-axis). From the left to the right: (first graph) the full mixed posterior deviance, (second graph) the deviance for the vegetation level (V, 3rd term in Eq. 3.5), (third graph) the deviance for the accumulation level (X, 2nd term in Eq. 3.5) and (fourth graph) the deviance for the data level (Y, 1st term in Eq. 3.5).

j	Carpinus	Corylus	Fagus	Quercus Ever	Quercus Dec	Grass & Shrubs
$b^j S^j_{i_1}$	0.19	5.07	20.21	58.3	66.24	9.11
$X^j_{i_1}$	0	1	1	25	2.03	3
$p^j_{i_1}$	0	0.03	0.03	0.69	0.06	0.08
$Y^j_{i_1}/(\Sigma_j Y^j_{i_1})$	0	0.001	0.006	0.85	0.06	0.05

Table 3.2: Value of the data and means of posterior quantities for the major taxa (in columns) of a point i_1. (first line) modelled pollen brought by dispersal to site i_1 for the taxa j. (second line) modelled pollen accumulated at site i_1. (third line) modelled proportions (i.e. *in situ* normalisation of the preceding line) of pollen accumulated. (fourth line) proportions estimated from the count data by normalisation.

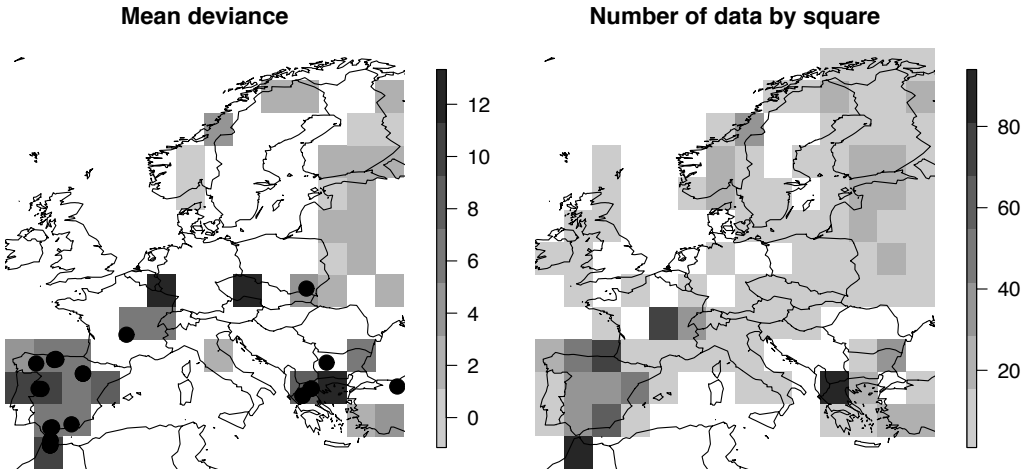

Figure 3.3: Spatial repartition of the difference between the mean of the deviance for the posterior and reference distributions. (left panel) Each square's colour represents the difference for the data points covered by the square. The black dots show the points whose difference is higher than 50. Squares covering less than 10 points have been removed for their unreliability. (right panel) Number of data points by square.

dispersed than specified by the multinomial distribution; what is indicated by the posterior predictive discrepancies. But in the case of the pollen brought by dispersion and accumulated ($b^j S^j_{i_1}$ versus $X^j_{i_1}$) it is evident that the Poisson distribution is not adequate. Indeed, for the deciduous Quercus (oak), cumulated probability above 3 for a Poisson distribution centred on 66.24 is around 2.2 10^{-21}. This high $b^j S^j_{i_1}$ value of pollen theoretically brought by dispersal is due to the nearby points which show high pollen percentages for this taxa. This is an argument in favour of more overdispersion than specified by the Poisson distribution at this level.

The second potential source of data-model conflict is the hypothesis of spatial stationarity of all the parameters. Indeed, we have the experience that LPJ-GUESS vegetation model used to simulate the NPP makes generally stronger errors in southern Europe than in northern Europe (see discussion about precipitation in Garreta et al., 2009). To check this aspect of the problem we plot the spatial repartition of the difference between posterior and reference deviance for each point over Europe (Fig. 3.3). As

Taxa (j)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
q^j (in %)	58	53	87	70	90	67	74	84	54	54	53	86	82	65	60
m^j/E^j	0.32	0.52	0.82	0.17	0.85	0.58	0.9	0.56	0.42	0.25	0.56	0.76	0.79	1.34	1.51
$\sqrt{\sigma^j}/E^j$	2.0	1.8	2.4	1.6	1.4	2.1	1.2	1.8	1.9	1.6	1.8	1.1	1.2	0.5	3.9
γ^j (in km)	53	90	84	154	78	30	193	58	80	20	30	57	224	448	50
b^j/b^{15}	2.03	0.64	0.84	0.22	0.24	0.06	0.05	0.08	0.44	0.05	0.08	0.03	0.04	0.01	1

Table 3.3: Table of the posterior mean of the parameters (for taxa names see Table 3.1). E^j is the mean of NPP^j when $\text{NPP}^j > 0$. q, m and σ are interpreted in term of anthropisation and/or vegetation model's error. γ is a dispersal length parameter and b is related to the relative pollen production between species and the overdispersion of the multinomial distribution.

expected, the deviance is higher in the South (mainly is Spain and Greece) and Centre (France, Switzerland, Belgium and Austria) than in the North. For the southern points we interpret these high discrepancies by a lack of realism of the model. For the central points this may be due to an anthropogenic disturbance significantly higher than for the other points, which cannot be accounted for since the parameters (m, σ and q) controlling such disturbance are constant over Europe. We come back on this stationarity problem in the discussion.

Output summaries

MCMC chain outputs give information about mixing and convergence. The number of parameter chains is $k * 5 = 60$. We show those related to hazel trees (Corylus) in Figure 3.4. Recall that plotted iterations are one iteration over 100 and after the iteration 400k; iterations for which the algorithm seems to have reached convergence.

We give the posterior mean of all the parameters Table 3.3. The q, m and σ parameters are related to the anthropogenic disturbances and/or the errors in the vegetation model simulations. One trait of these posteriors is that for most of the taxa (13 over 15), m^j, the mean of $V_i^j > 0$ despite that $\text{NPP}_i^j = 0$ is lower than the mean of $\text{NPP}^j > 0$.

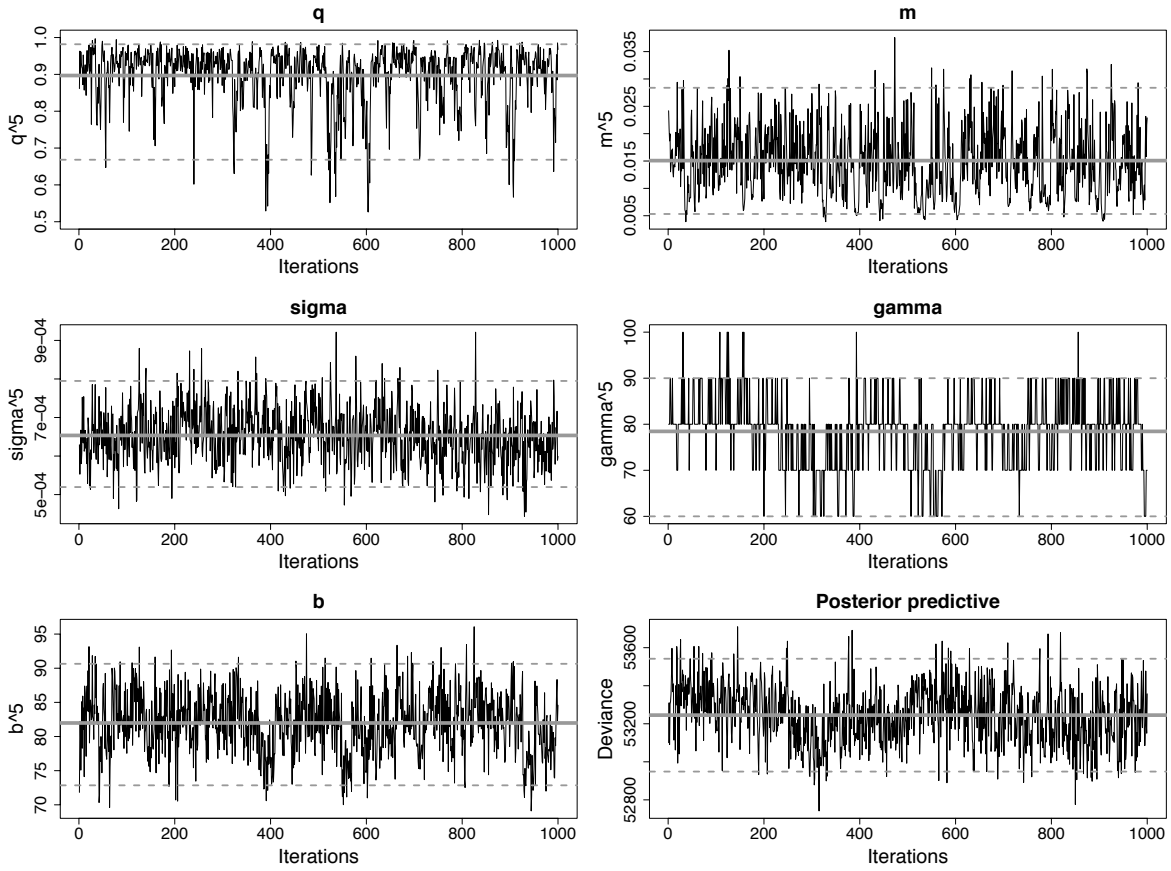

Figure 3.4: Sequence of posterior simulations of the parameters of the Corylus taxa (hazel tree) and variation of the posterior predictive deviance over the iterations (graph at the bottom right). Iterations plotted are one iteration over 100 between iteration 400k and 500k. The light grey lines show the mean (solid line) and borders of the 95% HPR (dashed lines).

This is expected: when the vegetation NPP has not been simulated, V should not be more productive than in places where NPP has been simulated.

The posterior values of the spatial dispersal parameters (γ^j, Table 3.3) range from 20 up to 220 km except Populus taxa (14, Poplar) which has a range of 450 km. This high range may result from a poor prediction of the species presence by the vegetation model. Indeed, the species is present almost everywhere along rivers and the model does not take rivers into account. In case of poor spatial matching between model simulations and data, the statistical model may need the maximum of dispersal to link them. The range of the other dispersal parameters is one order of magnitude higher than the length of a pollen flight between the canopy and the ground. This highlights that we do not look at the same processes that in individual tree based studies. Our modelling of vegetation is in term of populations instead of individuals and the scale of sampling is Europe. Then, γ dispersal parameters integrate a homogeneity in the vegetation (when a taxa is present at a point it is often present several km around) plus a classical dispersal term. The dispersal component produces very smooth fields of dispersed pollen $b^j S^j$. The one of the Corylus is presented in Figure 3.5.

The parameters b^j have two major interpretations. First, they are related to the relative pollen production between species. The values b^j/b^{15} (Table 3.3) is the relative pollen production between the taxa j and the 15th (Grass/Shrubs group). Second, the absolute values b^j contribute to the overdispersion of the multinomial which is easily measured through the term $K = \Sigma_{j=1..k}\ b^j S^j$. The posterior mean of this term is 38 with a 95% HPR equal to $[7; 70]$. Since this value is finite and well constrained, a non-negligible overdispersion of the multinomial distribution is inferred. We saw in the preceding validation step that this overdispersion is underestimated, at least, at the Poisson level. Thus, the question is, why is this K parameter so high? This certainly results from the set of intrinsic constraints imposed to the b^j parameters. Recall that they drive relative production between species and overdispersion through the Poisson distribution centred on $b^j S^j$. This means that when a low productive taxa j is present at a site i, its associated term b^j, relatively lower than the others, has to be sufficiently high to allow the generation of, at least, a $X_i^j = 1$ from a Poisson distribution centred

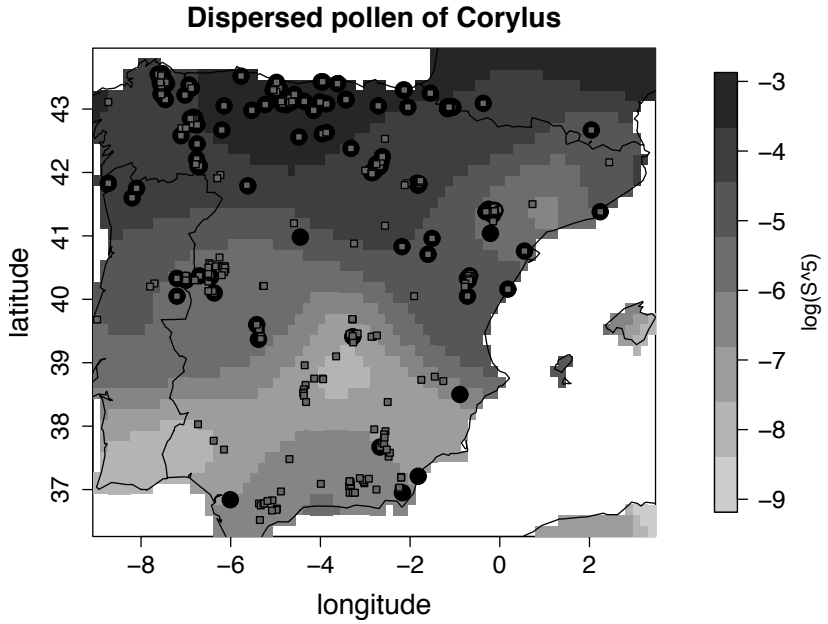

Figure 3.5: Posterior mean of the dispersed pollen of Corylus (Hazel tree, S^5) over Spain. Black dots show sites where Corylus pollen is present (i.e. in the data $> 0\%$) and squares show sites where Corylus presence has been simulated by the vegetation model. The scale is logarithmic.

on $b^j S_i^j$. This constraint reduces the range of possible b (and therefore K) values. We discuss alternative modelling of the overdispersion in the following section.

3.5 Discussion

In this paper, we proposed a statistical model for the so-called calibration of a transfer function (TF) linking simulated vegetation using a vegetation model and pollen records. Our TF is entirely process-based and parametrical. Such process-based model is preferable to classic TF because (a) it can be better evaluated, e.g. the parameter values are comparable to known quantities and the modelled processes are partially known, (b) it theoretically leads to more realistic parameters and uncertainty estimates (one of the critical issues when reconstructing past climate, IPCC Core Writing Team, 2007), and (c) because the modelling hypotheses are based on processes, their rejection or acceptance provides an information which is useful and communicable outside the community of statisticians for the palaeoclimatology. In this sense, this work brings together close disciplines: palynology, ecology, vegetation modelling, climatology and creates interactions between them. Nevertheless this approach has the flaw of needing a lot of information and heavy computing time. This makes its inversion for palaeoclimatological reconstructions an open and challenging problem in statistics.

There are two major underpinnings in our approach, first, the spatial stationarity principle which is fundamental to most data-based palaeoclimatology approaches and, second, the need to model the very evident spatial correlation in the pollen samples due to pollen dispersion (neglected or too simply approximated in classic TF).

The spatial stationarity principle is explicit in our approach instead of being hidden or implicit in classic TF. It is a corollary of the palaeoclimatology principle: 'use modern spatial variations to quantify past temporal variations'. Indeed, if the model calibrated on the modern - spatial - dataset is not stationary in space, i.e. if it includes variations in its parameters depending directly on longitude and latitude, it becomes impossible to invert it in time. In other words, by fitting the model to space, one admits that spatial variations in data are not generated by a general process that can be used to calibrate a general statistical model. When apparent, the non-stationarities in space must be modelled as coming from spatialised variables (if we suspect that these variables have

an impact) or properly quantified as a, potentially spatial, error.

Pollen dispersal along with spatial correlation in the distribution of plants are the major processes generating the spatial correlation in pollen surface samples. To model it, we used a Gaussian dispersal kernel (copying the Sutton's equation used in palaeoecology) which is convoluted over the mixture model for the vegetation random field (V). This approach is the Bayesian hierarchical 'process convolution' or 'moving average' approach of Higdon (1998) applied, in our case, to a non-Gaussian random field. It is interesting to remark that this approach, which was initially created to build spatially structured random fields (Barry and Ver Hoef, 1996) without explicit connection to a natural process is here based on an intuitive dispersion process. Several ways of reducing its computational burden have been proposed. The main one, proposed by Higdon (1998), consists in reducing the dimension of the problem by simulating the vegetation over a coarse grid instead of at each pollen points. This implies to select a grid size and requires proper mathematical studying since it changes the initial homotopic inference problem (pollen and vegetation at the same location) to an heterotopic one.

3.5.1 Over-dispersion and zero-inflation of the multinomial

Since we decided to avoid the physically based modelling of the accumulation process, we had to select one model for the overdispersion of the multinomial. We proposed the MP model, which was rejected by all the model checking criteria. We come back on the need for an overdispersed and zero-inflated multinomial model and propose ways to extend the MP model for our problem.

The classic model for multinomial overdispersion is the Multinomial Dirichlet model (MD, Leonard, 1977). We did not select it since it does not allow for zeros in the vector p; the 'structural zeros' coming from the absence of certain taxa over certain regions. We proposed the MP model because, with the same number of parameters than in the MD it provides two sources of zeros: the multinomial and the Poisson distribution. From another point of view, this model introduces a continuum in the zeros at the Pois-

son level, from those due to absence of the taxa $S_i^j = 0$ to those, less frequent, generated by little S_i^j (low vegetation values around the site i or vegetation far from the site). This is a vision opposed to the one represented by the mixture model for zero-inflation in which the extra-zeros are generated by a distinct process modelled through a Dirac mass at zero.

However, the MP model is not adequate to model pollen accumulation and sampling. This has been identified by the model checking criteria. We showed that this is likely due to a lack of overdispersion and/or zero-inflation. A simple way to overdisperse the MP model is to replace the Poisson distribution by a Negative Binomial. This can be done by setting:

$$[X_i^j | b^j, \tau^j, S_i^j] = \mathcal{NB}(b^j S_i^j, \tau^j)$$

with $\mathcal{NB}(x, m, s) = \frac{\Gamma(x+\tau)}{x!\Gamma(\tau)} \left(\frac{\tau}{m+\tau}\right)^\tau \left(\frac{m}{m+\tau}\right)^x$ the Negative Binomial distribution with mean m and variance $m + m^2/\tau$. This distribution has a set of k τ^j parameters more than the Poisson distribution. We cannot make a demonstration as we did for the Poisson but intuitively $k-1$ b^j parameters should control (and therefore be controlled by) the p^j. The k variance parameters τ^j may be identified on the variance of each p^j. It seems that the kth parameter b^j will have to be fixed if it does not account for discretisation.

People preferring continuous distributions and mixing would have the choice between distributions more dispersed than the $\mathcal{G}(b^j S_i^j, 1)$, such as $\mathcal{G}(b^j S_i^j, \tau^j)$, the continuous counterpart of the Negative Binomial, or mixing between gamma and Dirac mass at zero. The choice is the same as between Poisson (MP) and Gamma (MD) distributions: do we need one more source of zeros than the multinomial? In the case of mixing of distributions compared to the Negative Binomial, do we need a continuum between the process generating zeros and the one generating positive values?

3.5.2 Palaeoclimatology, other palaeo-sciences and vegetation model inversion

The core of our model - the representation of the pollen production and dispersion and the multinomial distribution for sampled pollen - builds a conceptual bridge between paleoclimatology and ecology. Indeed, an analogous core structure is studied in a long series of papers starting with Parsons and Prentice (1981) and reaching its latest expression with the models of Sugita (2007a) and Paciorek and McLachlan (2009) (e.g. the review of Broström et al., 2008). These papers study the modelling of dispersion and accumulation at a local level using pollen samples in lakes and vegetation records several km^2 around. One of their major research questions is palaeo-landscape reconstruction as an indicator of human constraint. Human constraint or disturbance is clearly the process which needs our attention. Indeed, the interpretation of our parameters (m, q and σ), the stationarity problem related to them and in a more general vision of the vegetation model, its calibration and validation *cannot* go ahead for a long time without considering an effect of human on the vegetation. In a climate reconstruction perspective, the modelling of the human disturbance and its quantification are also problematic. When inverting the climate-vegetation-pollen relation based on past pollen records, humankind has an influence on this relation which varies in time (e.g. St. Jacques et al., 2008). In our framework, if we interpret the m, q and σ parameter as pure nuisance parameters due to anthropisation, they should be set to 0 when reconstructing climate of periods at which humankind was not present.

Climate reconstruction by the model inversion is very promising but it is a great challenge. Indeed, combining our spatial representation of the pollen/vegetation link and the dynamical (over time) inversion of the vegetation model presented in Garreta et al. (2009) would be a great step toward a unified and fully process-based spatio-temporal method for climate reconstruction. This would require a great amount of work, at least, to build efficient spatio-temporal inversion algorithms combining particle filtering and MCMC and to consider validation in such a computationally over-intensive task (for

model validation in computationally intensive inverse problems, see e.g. Bhattacharya and Haslett, 2004). This would also need to consider the modelling of several sources of uncertainties and correlations attached to the temporal context, mainly the dating uncertainties and the temporal correlation in climate (for a discussion of these points, see Haslett et al., 2006).

Acknowledgements

The first author would like to thank Paul Miller for support with LPJ-GUESS. A first version of the demonstration in the appendix has been greatly improved by this version based on the Binomial distribution which was proposed by Professor John Haslett (TCD, Dublin). This work has been funded by the European Science Foundation (ESF) under the EUROCORES Programme EuroCLIMATE (project DECVEG) and by the French Centre National de la Recherche Scientifique (CNRS).

Chapter 4

Bayesian semi-mechanistic modelling for a process-based palaeoclimatology

This chapter is an article prepared for a submission to an applied statistical journal. Part of the reflection comes from discussions around a postdoc project we wrote and submitted with Professor John Haslett (Trinity College, Dublin, Ireland).

Abstract Pollen-based palaeoclimate reconstructions are based on models of the relation between the environment controlling plant species and pollen collected in sediments. Most of these models are descriptive and based on the same irreducible set of hypotheses. Their dependence on the same - not testable - set of hypotheses cuts into one's confidence in reconstructions of past climate that are all based on these methods. A process-based approach forms a necessary complement to correlative models by being based on different hypotheses supported by modern research in ecology. The combination of a mechanistic (vegetation) model and stochastic modelling is crucial to achieve the process-based modelling of the complex environment-plant-pollen system by (i) capturing the up-to-date knowledge in ecology included in the computer model and, (ii) properly quantifying and accounting for the various sources of uncertainties in the system through statistical modelling.

We propose the coupling of a Dynamic Vegetation Model (LPJ-GUESS) for the environment-plant relation and a statistical hierarchical model for the plant-pollen relation. The Bayesian paradigm allows us to consistently embed the computer simulator into the statistical model. The main challenge is the inference of such a composite model due to the computing cost of simulating from LPJ-GUESS and the large number of data linked by spatio-temporal relations. We propose a first approach for inference using Monte Carlo methods.

We present and apply our approach into the two-step process of palaeoclimatology. First, the *calibration* of the statistical model is realised using a spatial dataset of climate and pollen samples from Europe. Second, the *reconstruction* of past climate dynamics are performed for four cores covering the Holocene, located in southern Sweden.

Reconstruction results are coherent with their pollen records. They show constrained changes for temperature but lack a strong constraint for precipitation. The large differences between sites show the need to account for the processes of spatial vegetation dynamics (e.g. migrations). This promising method still requires theoretical and technical works in statistics to bypass the approximations we necessarily used and readily allow (a) the calibration of parameters inside the DVM and (b) the spatio-temporal reconstruction of paleo climate and vegetation from several cores at the same time.

4.1 Introduction

Pollen-based palaeoclimate reconstructions are based on models of the relation between the environment controlling plant species and pollen collected in sediments. Except in the 'model inversion' method (Guiot et al., 2000), all these models, called Transfer Functions (TF), are purely statistical descriptions of the link between a few climate variables and the pollen assemblages collected in sediments. These classical TF include, for example, the Indicator Species (IS, Iversen, 1944), the Modern Analogue Technique (MAT, Hutson, 1980), the Response Surface (RS, Bartlein et al., 1986), the Weighted Average-Partial Least Square (WA-PLS, ter Braak et al., 1993), the Probability Density Function (PDF, Kühl et al., 2002) and a semi-parametric Bayesian approach of the RS (Haslett et al., 2006). Their lack of process modelling makes them *correlative* TF, based on the same two hypotheses. First, they assume that a few (from 2 to less than 10) climatic variables drive the plant species presence/absence or abundance. Second, they assume an 'instantaneous' species response to climate, i.e. a nearly constant equilibrium between species and climate allowing independent (in space and time) calibration and use of the TF. These hypotheses simplify the modelling and inference, which, however, remains a statistical challenge (Haslett et al., 2006) but they seem rather restrictive while models of the vegetation dynamics exist (e.g. reviews of vegetation models in Prentice et al., 2007). Up-to-date vegetation models include a more elaborated description of the species requirements than just climate variables (e.g. atmospheric CO_2 concentration, soil description) and simulate vegetation dynamics. Ecology thus provides the theoretical background, the tools and the arguments to start the building of process-based TF not requiring both hypotheses mentioned before. These TF would provide past climate reconstructions independent from those of the correlative TF, increasing our confidence in past climate reconstructions that are only available through these methods.

As proposed by Guiot et al. (2000), including a vegetation model in the TF allows one to readily account for part of the processes forming the environment-plant-pollen

relation. Here, we propose the first full process-based approach by coupling a Dynamic Vegetation Model (DVM, e.g. Prentice et al., 2007) for the environment-plant relation and a statistical, process-based, model for the plant-pollen relation. LPJ-GUESS (Smith et al., 2001) is a DVM simulating *stochastic* vegetation *dynamics* from monthly climate, CO_2 and soil descriptions. For instance, from a Net Primary Production (NPP) of several plant species at time $t-1$, say NPP_{t-1}, and CO_2, soil and climate chronologies linking times $t-1$ and t (C_t), it simulates a NPP at time t: NPP_t. Following Garreta et al. (2009) we then consider LPJ-GUESS as the conditional distribution

$$p_{\text{LPJ}}(\text{NPP}_t|\text{NPP}_{t-1}, C_t) \tag{4.1}$$

This distribution defined by LPJ-GUESS can be simulated from (by running the model) but cannot be evaluated for any given values of $(\text{NPP}_t, \text{NPP}_{t-1}, C_t)$ because it is defined through a complex chain of simulating mechanisms. It also defines a Markov transition distribution for the vegetation in time since it has the Markov property of depending only on the vegetation at the previous time step.

The vegetation model is coupled to a statistical model $p(Y|\text{NPP}, \theta)$ controlled by parameters θ, and representing the chained processes expected to link the NPP to the pollen sampled in sediments Y: local vegetation disturbance (error from the DVM) $\rightarrow$ pollen production $\rightarrow$ pollen dispersal $\rightarrow$ pollen accumulation $\rightarrow$ pollen sampling. The core structure of our model is then

$$p(Y_t, \text{NPP}_t|\text{NPP}_{t-1}, C_t, \theta) = p_{\text{LPJ}}(\text{NPP}_t|\text{NPP}_{t-1}, C_t)\, p(Y_t|\text{NPP}_t, \theta) \tag{4.2}$$

We call such coupling between a mechanistic (computer) model and a stochastic model, 'semi-mechanistic' modelling. It forms a powerful approach for obtaining process-based models of complex systems because (i) it incorporates the mechanistic processes contained in the vegetation model, (ii) it allows the proper quantification of noises in the relationship and, (iii) it provides us with a statistical framework for inference. Equations describing the processes forming a DVM are subject to intensive researches in plant ecophysiology. They are designed, calibrated and tested on various conditions and datasets (see the description of vegetation models, e.g. Prentice et al.,

1992; Smith et al., 2001). These datasets - partly expressed in the vegetation model - complement and enrich the information contained in the modern pollen and climate datasets used in palaeoclimatology. A second argument for the use of mechanistic models inside a TF is that their constant evolution and improvement (e.g. Prentice et al., 2007) can be incorporated into TF for a negligible cost. Moreover, the statistical component of a semi-mechanistic approach allows one to properly account for the errors, seen as the discrepancy between the mathematical (composite) model versus the 'real', not measurable, relationship expressed in the modern dataset. This approach also finally provides a framework and basic tools for inference, either of the parameters of the statistical model or the past climate state by inversion of the composite model based on pollen data.

The Bayesian paradigm is unequalled for the inference of semi-mechanistic models. For our application this is supported by theoretical and technical arguments. On the theoretical side, the causative nature of the process-based model imposes use of a relationship (Equation 4.2) in which pollen is a function of climate. Climate is then a fixed 'regressor' when calibrating the TF on modern dataset. The palaeoclimate reconstruction involves the inversion of the relation for reconstruction. This inversion, along with proper propagation of the errors from calibration to reconstruction is well defined in the Bayesian framework through a prior/posterior rationale as we will demonstrate in the next sections. In other frameworks for inference, e.g. Maximum Likelihood, this inversion is provided through heuristic algorithms without theoretical support. See for example, non-Bayesian reconstructions from response surfaces models (Bartlein et al., 1986; Gonzales et al., 2009) compared to Bayesian approach of these models (Haslett et al., 2006; Vasko et al., 2000). On the technical side, Bayesian inference tools can cope with the stochastic DVM defining a distribution only available through simulation. Indeed, inference algorithms based on Importance Sampling (IS, Robert and Casella, 1999) only requires simulations following the DVM distribution.

In the first section, we define the calibration of this process-based TF in a Bayesian

framework and overview the statistical model linking vegetation and pollen. In the second section, we define the process-based palaeoclimate reconstruction and discuss the strategy for the posterior distribution computation using available Monte Carlo tools. In the third section, we present palaeoclimate reconstructions for the Holocene at four sites in South Sweden.

4.2 A spatial TF for calibration

The first step of palaeoclimate reconstruction consists in inferring the parameters θ of the statistical model $p(Y|\text{NPP}, \theta)$ based on a set of N modern climate and pollen measurements. Given pollen surface sample collected at site $s = 1..N$, Y_s are multinomial vectors of counts per taxa $j = 1..k$, i.e. $Y_s = (Y_s^1, .., Y_s^k)$. In this study $k = 15$ pollen groups, see previous chapters for the grouping we made based on the hundreds of initial taxa. Modern climate conditions for the pollen sites, C_s, are monthly chronologies of precipitation, temperature and cloudiness. The sites are spread at a continental scale (Europe in our application) and the pollen surface samples are expected to show a spatial autocorrelation, at least due to the process of pollen dispersal and the spatial vegetation autocorrelation. We account for such autocorrelation in the statistical model linking vegetation simulated by the DVM and pollen. In the Direct Acyclic Graph (DAG) presented on Figure 4.1, we report the dependencies between measured climate, simulated vegetation and observed pollen surface samples.

Calibration is defined in a Bayesian framework as the posterior distribution of the parameters θ defining the climate-pollen link given the set of modern pollen and climate data. For convenience, in the following, the complete collection of pollen samples $y_{s=1..N}$ is noted $\mathbf{y}$, climates $\mathbf{c}$ and the modern vegetation simulations from LPJ-GUESS at all sites are noted $\mathbf{NPP}$. Following the model structure given by the DAG (Figure 4.1)

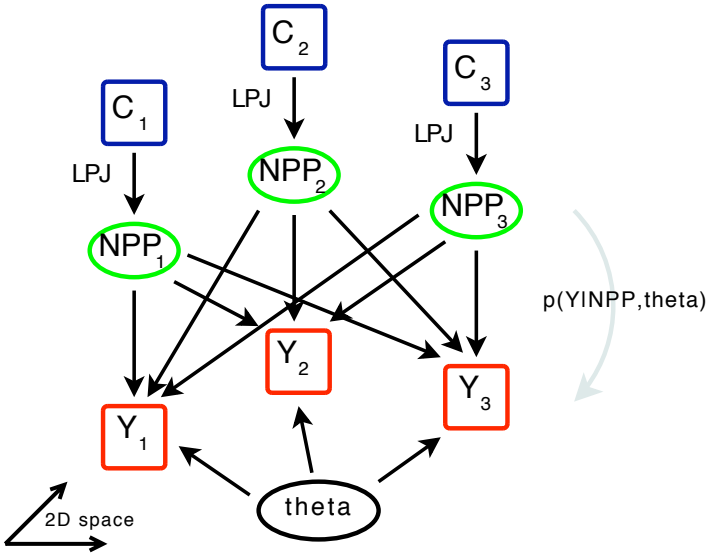

Figure 4.1: Directed Acyclic Graph (DAG) of the model used for calibration. Variables in squares, climate (C_s) and pollen (Y_s) are measured over Europe (at the point $s = 1..3$). The vegetation NPP_s, featured in a circle is not measured, i.e. hidden. It is simulated running the DVM LPJ-GUESS (Smith et al., 2001) for a given climate. The link between all the vegetation and pollen sites is a statistical model, noted $p(\mathbf{Y}|\mathbf{NPP}, \theta)$, with parameters $\theta = (\theta_1, \theta_2)$. The inter-relation between all the pollen samples and the vegetation simulated for all sites arises from pollen dispersal modelling in the statistical model.

and Bayes rule, the posterior distribution for the parameter calibration is

$$\begin{aligned} p(\theta|\mathbf{y}, \mathbf{c}) &= \int p(\theta, \mathbf{NPP}|\mathbf{y}, \mathbf{c})\, d\mathbf{NPP} \\ &= \int \frac{p(\theta, \mathbf{NPP}, \mathbf{y}|\mathbf{c})}{p(\mathbf{y}|\mathbf{c})}\, d\mathbf{NPP} \\ &\propto \int \left(\prod_{s=1}^{n} p_{\text{LPJ}}(\text{NPP}_s|c_s) \right) p(\mathbf{y}|\mathbf{NPP}, \theta)\, p(\theta)\, d\mathbf{NPP} \end{aligned} \tag{4.3}$$

with $p_{\text{LPJ}}(\text{NPP}_s|c_s)$ the distribution of the modern vegetation given 20th century climate. Ideally, to mimic the distribution in Equation 4.1, this distribution would be dependent on the vegetation at the beginning of the 20th century, which is poorly known. To use an equilibrium hypothesis between climate and vegetation that is as

weak as possible, we define $p_{\text{LPJ}}(\text{NPP}_s|c_s)$ by (i) spinning up the DVM with CO_2 concentration, abiotic, and climatic conditions of the beginning of the century (1901-1930) until equilibrium is reached, then (ii) simulating the whole century under the CO_2 and climatic conditions available for the century. NPP_s is defined as the mean of the DVM outputs between the years 1961 and 1990, which is expected to correspond with the time slice recorded by most of the modern pollen samples.

Obtaining the posterior distribution defined in Equation 4.3 is not realistic since this - necessarily - numerical integration, due to the implicit definition of $p_{\text{LPJ}}(\text{NPP}_s|c_s)$ through the computer simulator, is of dimension $\dim(\mathbf{NPP}) = N * k$ ($> 15,000$ in our case) and because it requires simulations from the DVM. For instance, a single simulation of NPP for the $N = 1301$ European points (one **NPP**) represents hours of computing.

We propose to calibrate directly the relation using pollen and a set of $\text{NPP}_{s=1..n}$ simulated under modern climate (called **npp**), and thus, to obtain

$$\begin{aligned} p(\theta|\mathbf{y}, \mathbf{npp}, \mathbf{c}) &= p(\theta|\mathbf{y}, \mathbf{npp}) \\ &= \frac{p(\theta, \mathbf{y}|\mathbf{npp})}{p(\mathbf{y}|\mathbf{npp})} \propto p(\mathbf{y}|\mathbf{npp}, \theta)\, p(\theta) \end{aligned} \tag{4.4}$$

By bypassing the DVM proper integration, we ignore the calibration of parameters 'inside' the DVM, i.e. the DVM calibration. This aspect is postponed to the discussion section. In the next sections we present an overview of the process-based statistical model linking the NPP_s and pollen samples Y_s in space. A complete description is given in the previous chapter of this thesis. We finally discuss inferences which, without DVM integration, remain difficult to obtain, due to the number of dimensions considered.

4.2.1 Process-based modelling of the vegetation-pollen link

Hierarchical (or conditional) modelling provides us with a simple framework for modelling a causative chain of processes. We then model the major processes linking vegetation and pollen through two hidden levels defining $p(\mathbf{Y}|\mathbf{NPP}, \theta)$, with $\theta = (\theta_1, \theta_2)$,

through the integral

$$p(\mathbf{Y}|\mathbf{NPP},\theta) = \int p(\mathbf{Y}|\mathbf{X})\, p(\mathbf{X}|\mathbf{V},\theta_2)\, p(\mathbf{V}|\mathbf{NPP},\theta_1)\; d\mathbf{X}\, d\mathbf{V} \tag{4.5}$$

with $\mathbf{V} = V_{s=1..N}^{j=1..k}$ a set of k latent, 'actual vegetation', fields sampled at the N sites and $\mathbf{X} = X_{s=1..N}^{j=1..k}$ a set of k latent, 'accumulated pollen', fields sampled at the N sites.

Here, attention must be paid to choosing X and V distributions that make the integral analytically tractable or, at least that define analytically full-conditional distributions for several variables. Obtaining an analytically tractable integral reduces too much the range of available models and was not considered. When possible without loss of physical realism, we chose distributions that are conjugated, i.e. allowing steps of Gibbs algorithm in the inference using Markov Chain Monte Carlo algorithms (Robert and Casella, 1999).

Hidden levels and processes

• Vegetation simulated by LPJ-GUESS is *potential*, i.e. controlled by climate, soil properties, CO_2 and not disturbed by human activities. Then, 'actual' vegetation composition for the site s, noted $V_s = (V_s^1, .., V_s^k)$, is expected to be a noisy image of the simulated NPP_s and modelled using a mixture of Dirac and Gamma distributions (more details in the previous chapter). The vegetation disturbances are modelled using independent distributions per site and taxa

$$p(\mathbf{V}|\mathbf{NPP},\theta_1) = \prod_{s=1}^{n}\prod_{j=1}^{k} p(V_s^j|\text{NPP}_s^j,\theta_1)$$

• Pollen production of the taxa j is linearly related to the actual vegetation field $\mathbf{V}^j$ through a parameter b^j and airborne pollen dispersal is modelled using a Gaussian dispersal kernel of parameter γ^j by species. Then, for any site s and species j, pollen transported by dispersal, $b^j S_s^j$, is the - deterministic - convolution of a Gaussian kernel over the vegetation spatial field V^j.

• Pollen accumulation in natural traps such as lake and mires has a major influence on the signal registered. This process is highly complex and depends on site specific variables (lake and basin sizes and shapes, wind direction, etc) that are poorly known. Therefore, we model accumulation as a random process centred on the pollen - theoretically - brought by dispersal ($b^j S_s^j$). From another point of view, since it is the hidden level conditioning multinomial sampling, it controls the multinomial overdispersion. Its choice is discussed in the next section. A general assumption is that, conditional on the vegetation fields (V^j), the distributions are independent among sites and taxa

$$p(\mathbf{X}|\mathbf{V}, \theta_2) = \prod_{s=1}^{n} \prod_{j=1}^{k} p(X_s^j | \mathbf{V}^j, \theta_2^j)$$

• A fraction of pollen accumulated is sampled, recognised and counted until a prespecified number N_s of pollen grains is reached. We model this as a independent multinomial distribution per site s with total outcome N_s and probabilities equal to the proportions of pollen j accumulated: $\forall\, j\ p_s^j = X_s^j / \Sigma_{j=1}^k X_s^j$.

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{s=1}^{n} \mathcal{M}(Y_s | p_s, N_s)$$

4.2.2 Multinomial overdispersion and zero-inflation

In the context of hierarchical modelling for multinomial data, the modelling of interrelated overdispersion and zero-inflation is achieved by defining the multinomial distribution conditional on probabilities p^j, $j = 1..k$ that are themselves randomly distributed. From the NPP_s to the dispersed pollen, our model is a spatial regression posed in terms of absolute quantities. Overdispersion and zero-inflation of the multinomial (pollen) data are modelled as coming from the process of pollen accumulation in the natural traps (peat bogs, lakes of different sizes, etc). For a lake, this process is roughly seen as the capture of airborne pollen over the area of the lake (Prentice, 1985). The lake is seen as a big urn in which captured pollen grains are shuffled before (multinomial) sampling is performed.

The three original models discussed in this section are based on the accumulation processes, suspected to create overdispersion and zero-inflation with respect to the Multinomial distribution. These process-based constructions are in the spirit of Ancelet et al. (2009), and differ from 'two-part' or 'hurdle' models (reviewed in Ridout et al., 1998; Martin et al., 2005) which assume different sources for the zeros on one side and the positive values on another. This makes our models parsimonious and close to threshold models (preceding references and Salter-Townshend and Haslett, 2006, for multinomial overdispersion), with a threshold fixed at zero.

In the previous chapter, we proposed to model the accumulation or capture following an independent (per taxa) Poisson distribution centred on the pollen quantity theoretically dispersed

$$p(X_s^j|\mathbf{V}^j, \theta_2^j) = \mathcal{P}(b^j S_s^j) \tag{4.6}$$

The quantities X_s^j are then normalised to form probabilities for the multinomial. This model is interpreted as follows: if pollen grains are dispersed on the ground following a Poisson process of intensity $b^j S_s^j$ (in grains m^{-2}) and every lake (and trap) has the same size α (in m^2). Then, each lake receives X_s^j pollen grains following a Poisson process of intensity $\alpha b^j S_s^j$. The lake size α is not identifiable since absolute pollen productions b^j are not identifiable because most pollen data are proportions. In the previous chapter, we did not use the term α and demonstrated that $k-1$ parameters b^j are interpretable in terms of relative pollen productions between species and the kth parameter $K = \Sigma_{j=1}^k b^j$ plays the role of α, controlling the overdispersion that is related to the lakes' size. When lake size decreases, K decreases and overdispersion increases, i.e. the lake represents less and less well the pollen composition brought by dispersal. This discrete model allows for zeros coming from the accumulation process. The probabilities of zeros are given by the Poisson distribution and then, are directly dependent on the overdispersion.

Using predictive posterior measures (Gelman et al., 1996), in the previous chapter, we detected that this Poisson model is too limited in overdispersion. This can be

interpreted as due to the restriction to a unique lake size, unable to represent the range of possibilities (from peat bogs and mosses to large lakes). The lake size could be randomly modelled as T_s following

$$p(X_s, T_s|\mathbf{V}^j, \theta_2) = p(T_s) \prod_{j=1}^{k} p(X_s^j|\mathbf{V}^j, T_s, \theta_2)$$

This modelling introduces one more latent field of lake sizes T, which gives to the whole model a too complicated structure for inference due to the prior dependence between fields $j = 1..k$ through the lake size. Using independent gamma-distributed T_s^j per site and taxa provides conjugacy with the Poisson distribution. This makes $p(X_s^j|\mathbf{V}^j, \theta_2)$ independent negative binomial distributions

$$p(X_s^j|\mathbf{V}^j, \theta_2) = \mathcal{NB}\left(\tau^j, \tau^j/(\tau^j + b^j S_s^j)\right) \tag{4.7}$$

centred on $b^j S_s^j$, with variance $b^j S_s^j + (b^j S_s^j)^2/\tau^j$ and probability of $X_s^j = 0$ equal to $\left(\frac{\tau^j}{\tau^j + b^j S_s^j}\right)^{\tau^j}$. This model accounts for site-specific random variations (e.g. lake size, winds, etc) and other taxa-specific random variations (e.g. non-homogenous distributions of the taxa around the site, etc). But there is no proof for the identifiability of b^j and τ^j, $j = 1..k$. We made simulation tests (results not presented here) showing that the model becomes identifiable when overdispersion is strong ($\tau^j < 10$). For τ^j around or higher than 10, the inference algorithm diverges quickly to high values (several $\tau^j > 1000$). Without formal demonstration, this indicates that the model can be inferred for very overdispersed models and when the model is not identifiable, this is detectable on the posterior distributions.

4.2.3 Inference using Markov Chain Monte Carlo

The weak point of process-based modelling is that it requires one to infer models with complex and original structures for which inference guidelines are not readily available. For instance, our model is composed of two sets of $k = 15$ latent fields (V^j and X^j with $j = 1..k$) sampled at $N = 1301$ points, and five or six sets of k latent variables because $\dim(\theta_1) = 3 * k$ and $\dim(\theta_2) = 2 * k$ or $3 * k$ depend on the model selected

for overdispersion Eq. 4.6 or Eq. 4.7. In the Bayesian framework, we use Markov Chain Monte Carlo algorithms (MCMC, Robert and Casella, 1999) that are adaptable to nearly any model. We used an adaptive Metropolis-within-Gibbs algorithm, i.e. Metropolis steps are performed inside the main Gibbs loop trough full-conditional distributions and proposal variances are tuned during a burn-in period. The sequential nature of MCMC algorithms makes them very slow to obtain the desired posterior simulations. We used the structure of the model to parallelise computation for each iteration. The model for accumulation per taxa and site are chosen independently, implying that full conditional distributions of X_s are independent between sites, and $\mathbf{V}^j$, θ_1^j and θ_2^j are independent between fields. The only difficulty when parallelising is the need for a good parallel random number generator. We used the one of L'Ecuyer (1999).

Convergence assessment and model testing are challenging since the inference is so long that it is performed once and for all. Convergence has been checked using a subset of the simulation outputs and deviance monitoring. Model testing (or adequacy checking) cannot use the methods based on cross-validation since this require one to re-perform the inference several times. Combination of posterior predictive diagnostics (e.g. Gelman et al., 1996) and mixed replications (Marshall and Spiegelhalter, 2007) are used to create several diagnostics for testing the hierarchical model at its various levels (see previous chapter).

4.3 Reconstruction of the past vegetation and climate dynamics

We consider the inference of past vegetation and climate dynamics from a sequence of $n+1$ pollen samples along a sediment core. Each sample is assumed to be dated without uncertainty (discussed in the conclusion). We note Y_t the pollen sample associated with the date t varying from t_0 (oldest) to t_n (youngest). Hence, we note $y_{t_0:t_n}$ the sequence of pollen data, $NPP_{t_0:t_n}$ the corresponding, unknown, sequence of vegetation and $C_{t_0:t_n}$

the climate.

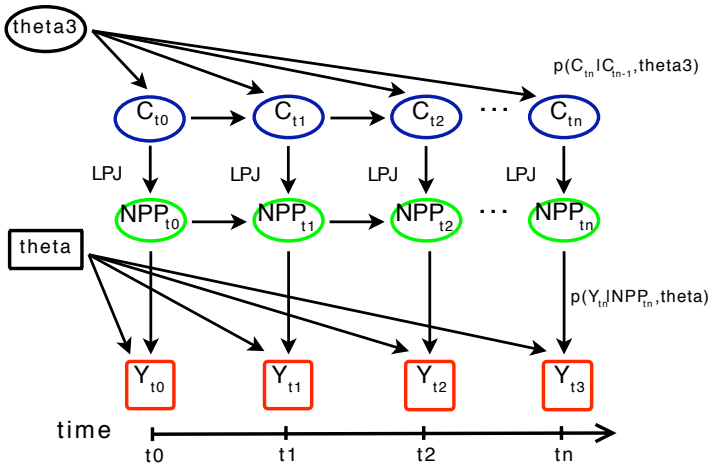

Figure 4.2: Directed Acyclic Graph of the model used for reconstruction. Variables in squares are known. Those featured in circles are to be reconstructed. The ages are given by the core points and range from t_0, for the oldest sample, to t_n, for the youngest one. Climate $C_{t_0:t_n}$ is a Markovian process indexed over time and driven by the unknown parameter θ_3. $\mathrm{NPP}_{t_0:t_n}$ is the vegetation defined by the DVM and $Y_{t_0:t_n}$ are the pollen samples. Parameter $\theta = (\theta_1, \theta_2)$ is known from calibration, i.e. distributed following $p(\theta|\mathbf{y}, \mathbf{npp})$.

Based on the structure of the previous section and the DAG presented in Figure 4.2, a Bayesian palaeoclimate reconstruction consists in finding the posterior distribution

$$\begin{aligned} &p(\mathrm{NPP}_{t_0:t_n}, C_{t_0:t_n}, \theta_3 | y_{t_0:t_n}) \\ &\qquad \propto p(C_{t_0:t_n}|\theta_3)\, p(\theta_3)\, p_{\mathrm{LPJ}}(\mathrm{NPP}_{t_0:t_n}|C_{t_0:t_n}) \prod_{t=t_0}^{t_n} p(y_t|\mathrm{NPP}_t) \end{aligned} \tag{4.8}$$

In the second line, the first term is a temporal model for climate, with parameters θ_3, followed by the prior for θ_3. We select and describe them later. The third term is the vegetation dynamics defined by the DVM through the conditioning chain (Eq. 4.1)

$$p(\mathrm{NPP}_{t_0:t_n}|C_{t_0:t_n}) = p_{\mathrm{LPJ}}(\mathrm{NPP}_{t_0}|C_{t_0}) \prod_{t=t_1}^{t_n} p(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_{t-1})$$

The fourth term is the product of $p(Y|\text{NPP})$ applied to the core points $y_{t_0:t_n}$. This distribution is defined as the model $p(Y|\text{NPP}, \theta)$ calibration previously whose parameter $\theta = (\theta_1, \theta_2)$ has been integrated out to account for the calibration uncertainties, i.e.

$$\begin{aligned} p(y_t|\text{NPP}_t) &= \int p(y_t|\text{NPP}_t, \theta)\, p(\theta|\mathbf{y}, \mathbf{npp})\, d\theta \\ &= \int p(y_t|X_t)\, p(X_t|V_t, \theta_2)\, p(V_t|\text{NPP}_t, \theta_1)\, p(\theta|\mathbf{y}, \mathbf{npp})\, dV_t\, dX_t\, d\theta \end{aligned} \tag{4.9}$$

In the first line, $p(y_t|\text{NPP}_t, \theta)$ is the model used in calibration (Equation 4.5) and $p(\theta|\mathbf{y}, \mathbf{npp})$ the posterior obtained in calibration, which carries uncertainties on θ and is available under the form of simulations. This is the Bayesian way for transferring uncertainties from the calibration to the reconstruction (Haslett et al., 2006). In the second line, we expand the model following the two hidden levels defining the pollen-vegetation link.

If the temporal structure of climate is chosen to be Markovian, the reconstruction model is a hierarchical and continuous hidden Markov model referred to as a *state-space* model (Cappé et al., 2005). This qualifies palaeoclimate reconstruction as the joint inference of hidden states (vegetation and climate at each time point) and parameter θ_3. Inference for such models is in itself an active field of research (e.g. Cappé et al., 2005; Andrieu et al., 2010, and references therein). Here, the model has the major originality of embedding a mechanistic (computer) model which implicitly defines its core structure. This precludes any approach requiring to compute the transition kernel $p(\text{NPP}_t|\text{NPP}_{t-1}, C_t)$ and changes the constraints on inference regarding computing time and memory size. Indeed, the computer simulator defines a *forward* transition kernel that can only be simulated from. The simulations from the DVM are the most time-demanding task for inference. Their length is proportional to the time range considered: $t_n - t_0$. For instance, simulating the vegetation over 12 kyr (thousands of years) represents 25 min of computing time. In the following, computing time will be expressed relative to the time range considered ($t_n - t_0$).

The main types of algorithms used for the inference of state-space models are MCMC

(Robert and Casella, 1999) and Sequential Monte Carlo (SMC or Particle Filter, Doucet et al., 2001). A MCMC type algorithm starts from a set of initial parameters and series of vegetation and climate for times t_0 to t_n and sequentially scans the posterior distribution by global or local moves. The huge number of dimensions and the absence of global conjugacy properties preclude the use of a Metropolis algorithm with a global proposition because these propositions will always be rejected. Although it cannot work, this type of algorithm would have an ideal computing cost of $t_n - t_0$ for a single proposition. Due to the computer model, a local move for climate and vegetation at time t requires to re-simulate the vegetation for the time points t to t_n. Indeed, if a new climate and vegetation transition from $t-1$ to t is accepted, it has to be linked to $t+1$ which is typically not possible since we can only simulate a new $t+1 : t_n$ series. Then, if time points are equally spaced on the core, the computation time associated to this strategy is approximately $(t_n - t_0) * n/2$ for a single global move (sum of local moves for all points). This is not realistic because n is often greater than one hundred. A trade-off can be found using 'regional' moves, i.e. moves of several points at the same time, but this is restricted by the need for reasonable acceptance rate and the cost is bounded between $(t_n - t_0) * n/2$ and $t_n - t_0$ for a single proposition. In any case, MCMC approaches will always face the problem that increasing the number of core points (n, i.e. the quality of data) increases computing time. This problem is serious because a future spatio-temporal approach, by merging cores over space, would increase the number of time points not aligned between cores.

A SMC-type algorithm is far more natural in the context of forward temporal simulations. Indeed, such algorithms are based on Importance Sampling (IS), i.e. DVM simulations and weighting, and treat the problem time after time. We will present a first approach of inference with a simple SMC algorithm that allows (a) 'smoothed' inference of the static parameters θ_3 and, (b) 'filtered' inference of the climate and vegetation states. Its computing time for a single 'particle' (corresponding to an independent move) is $t_n - t_0$. In the Bayesian framework, 'smoothed' and 'filtered' inferences respectively describe the inference of a quantity given all the data and the inference of a time-t quantity given data before and at time t (e.g. Cappé et al., 2005). The

ultimate goal is to obtain smoothed inference for climate and vegetation. We hope that this work will be the basis for developing a computationally tractable smoother.

4.3.1 Inference using a sequential Monte Carlo algorithm

In this section we present the general SMC algorithm. θ_3, the parameter of the climate structure is assumed to be known and not included as a parameter but it appears as a subscript when it is involved in a distribution. We come back on its inference in Section 4.3.1.

Following Doucet et al. (2001) and Andrieu et al. (2010), the inference of our state-space model (Equation 4.8) is achieved by sequentially approximating the sequence of distributions $p_{\theta_3}(\mathrm{NPP}_{t_0:t}, C_{t_0:t}|Y_{t_0:t})$, for $t = t_0..t_n$. This is obtained by exploiting the temporal structure of the model, rewritten in the recursive form

$$\begin{aligned} p_{\theta_3}(\mathrm{NPP}_{t_0:t}, C_{t_0:t}|Y_{t_0:t}) \propto p_{\theta_3}(\mathrm{NPP}_{t_0:t-1}, C_{t_0:t-1}|Y_{t_0:t-1})\, p_{\theta_3}(C_t|C_{t_0:t-1}) \\ p_{\mathrm{LPJ}}(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t)\, p(Y_t|\mathrm{NPP}_t) \end{aligned} \tag{4.10}$$

The algorithm works as follow: suppose at time $t-1$ we have a discrete approximation of $p_{\theta_3}(\mathrm{NPP}_{t_0:t-1}, C_{t_0:t-1}|Y_{t_0:t-1})$,

$$\hat{p}_{\theta_3}(\mathrm{NPP}_{t_0:t-1}, C_{t_0:t-1}|Y_{t_0:t-1}) = \sum_{m=1}^{N_p} \hat{\omega}_{t-1}^m \delta_{\{\mathrm{npp}_{t_0:t-1}^m, c_{t_0:t-1}^m\}}$$

where $\{\mathrm{npp}_{t_0:t-1}^m, c_{t_0:t-1}^m\}$, $m = 1..N_p$ are realisations of the corresponding random variables, called 'particles'. δ is the Dirac mass function and $\hat{\omega}_{t-1}^m$ are weights summing to 1. The time-t distribution (Equation 4.10) is obtained using the following pseudo code:

1. sample N_p random variables $(\mathrm{npp}_t^m, c_t^m)$, $m = 1..N_p$ following a proposal distribution $f_{t,\theta_3}(\mathrm{NPP}_t, C_t)$ and concatenate them to the preceding particle set $\{\mathrm{npp}_{t_0:t-1}^m, c_{t_0:t-1}^m\}$, forming the new particles set $\{\mathrm{npp}_{t_0:t}^m, c_{t_0:t}^m\}$,

2. compute the (non-normalised) importance weight of each particle

$$\omega_t^m = \hat{\omega}_{t-1}^m \frac{p_{\theta_3}(c_t^m|c_{t_0:t-1}^m)\, p_{\mathrm{LPJ}}(\mathrm{npp}_{t,s}^m|\mathrm{npp}_{t-1,s}^m, c_{t,s}^m)\, p(y_{t,s(t)}|\mathrm{npp}_{t,s_1:s_l}^m)}{f_{t,\theta_3}(\mathrm{npp}_t^m, c_t^m)} \tag{4.11}$$

3. normalise the weights so that their sum is 1

$$\hat{\omega}_t^r = \omega_t^r / \sum_{m=1}^{N_p} \omega_t^m \tag{4.12}$$

These steps provide a discrete approximation of the time-t distribution $\hat{p}_{\theta_3}(\mathrm{NPP}_{t_0:t}, C_{t_0:t}|Y_{t_0:t})$ under the form of weighted particles.

Classically, a *regeneration* step is included after step 3 if the series of weights $\hat{\omega}_t^m$, $m = 1..N_p$ degenerates too much, i.e. if non-null weights are carried by too few particles. Indeed, the recursive weighting (Equation 4.11) induces a degeneracy in the particle set which is prevented by sampling with replacement (i.e. duplicating or killing) the particles following their weights and setting all weights to $1/N_p$. We use the residual resampling procedure of Liu and Chen (1998). Resampling includes a Monte Carlo error in the posterior approximation and is performed only when degeneracy is too high; measured following that the Effective Sample Size criterion (ESS) is lower than $N_p/2$:

$$\mathrm{ESS}_t = \left(\sum_{m=1}^{N_p} (\hat{\omega}_t^2) \right)^{-1}$$

The ratio Equation 4.11 is intractable due to $p(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t)$ implicitly defined through the vegetation model. This problem may be bypassed by using the implicit distribution as part of the proposal distribution, i.e.

$$f_{t,\theta_3}(\mathrm{NPP}_t, C_t) = p_{\mathrm{LPJ}}(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t)\, g_{t,\theta_3}(C_t) \tag{4.13}$$

where $g_{t,\theta_3}()$ is a proposal distribution for C_t. This proposal distribution is clearly not optimal. Therefore, the use of a computer simulator defining an intractable distribution strongly constrains the search for an optimal proposal distribution (for this discussion, Doucet et al., 2001; Andrieu et al., 2010, and references therein). In our problem, this is reduced to the search for a 'good' $g_{t,\theta_3}(C_t)$. This search can be done using a mathematical rationale, e.g. seeking for a g proportional to the numerator Equation

4.11, then integrating out numerically the NPP etc. This is not realistic (integration requires re-running the DVM) and we use the following heuristic.

Using a fast non-parametrical TF (the Modern Analogue Technique, see application) we reconstruct the climate variables $C_{t_0:t_n}$ from several cores located around the site considered, concatenate the reconstructions and use their mean as the proposal (g_t) mean. The proposal g_t for the climate variables considered (see application) is then Gaussian with mean equal to the mean at time t and its variances and correlation are selected to be very large with no correlation to allow a wide range of variations around the mean values. The context of IS attached to the SMC algorithm provides a theoretical justification for using such approach, i.e. the reconstruction estimator will be unbiased and with finite variance until the proposal distribution has heavier tails than the (unknown) target (Robert and Casella, 1999). In this respect, the Gaussian distributions could be replaced by more heavily tailed distributions. The use of large variances for the proposal is more problematic since the IS scanning is limited to a small number of particles (here 1000). This produces poor representations of the posterior distributions that may occupy a small portions of the proposal support. To improve this representation we then used a two-pass reconstruction process 'in the spirit' of Cappé et al. (2004); Beaumont et al. (2010), but with less gain than in these approaches. This consists, in the first pass, in using the proposal discussed before and in the second pass we use the enlarged filtering posterior smoothed for each point of the first pass. Enlargement consists in doubling the posterior variance obtained at the first pass. The theoretical justification for using such two passes (and more if computing time allows it) arises as before from the IS nature of the SMC algorithm. Practical justification is that it allows memory size reduction, e.g. comparing a single pass with 2000 particles or two passes with 1000 particle and the reconstructions seem more accurate because they depend less on the prior used at the first pass. Evidently, precautions have to be taken. When the posterior variances seem unrealistically small (e.g. less than 1°C) they are enlarged. In the same idea, if the posterior distribution is located on an extreme part of proposal distribution this indicate a too restrictive proposal that may be improved sequentially by re-running the algorithm.

Then, using Equation 4.13, the ratio Equation 4.11 becomes

$$\omega_t^m = \hat{\omega}_{t-1}^m \frac{p_{\theta_3}(c_t^m | c_{t_0:t-1}^m)\, p(y_t | \mathrm{npp}_t^m)}{g_{t,\theta_3}(c_t^m)} \tag{4.14}$$

Computation of this ratio remains difficult due to the high dimensional integration needed for $p(Y_t|\mathrm{NPP}_t^m)$ (see Equation 4.9). In the next section we present an approach for solving this problem and the following section is dedicated to the modelling and inference of a temporal climate structure with parameter θ_3. In the final section, we discuss the inference of climate from several close cores at the same time.

Particle weighting by high dimensional integration

For each time point t (from t_0 to t_n) of a single core, the SMC algorithm provides N_p (=1000) vegetation particles npp_t^m whose weighting of each one requires computing the integral shown in Equation 4.9

$$p(y_t|\mathrm{npp}_t^m) = \int p(y_t|X_t)\, p(X_t|V_t,\theta_2)\, p(V_t|\mathrm{npp}_t^m,\theta_1)\, p(\theta_1,\theta_2|\mathbf{y},\mathbf{npp})\; dV_t\, dX_t\, d\theta_1\, d\theta_2 \tag{4.15}$$

Even in this only temporal case, the integral dimension is large (90 or 105 depending on the model for overdispersion Eq. 4.6 or Eq. 4.7) precluding the use of a 'brute force' IS integration per particle, which would require too many simulations. The number of integrals to perform is even more large ($N_p = 1000$) which makes too slow the use of a MCMC integration per particle npp^m. Since all these integrals have a common part due to the weighting based on the same y_t pollen data, we propose to produce a single 'omnibus' MCMC sample (X_t^r, V_t^r), $r = 1..M$ and use it for the weighting of all the particles through an IS scheme. We first present the integration for a single core before discussing the increasing dimension problem of integrating for several spatially distributed cores.

An IS estimate, $\hat{I}_t^m$, of the integral Eq. 4.15 is obtained by sampling $(X_t^r, V_t^r, \theta_1^r, \theta_2^r)$,

$r = 1..M$ following $p_{\mathrm{IS},t}(X_t, V_t, \theta_1, \theta_2)$ and computing

$$\hat{I}_t^m = \frac{1}{M} \sum_{r=1}^{M} \frac{p(y_t|X_t^r)\, p(X_t^r|V_t^r, \theta_2)\, p(V_t^r|\mathrm{npp}_t^m, \theta_1^r)\, p(\theta_1^r, \theta_2^r|\mathbf{y}, \mathbf{npp})}{p_{\mathrm{IS},t}(X_t^r, V_t^r, \theta_1^r, \theta_2^r)} \tag{4.16}$$

By theorem, this estimate converges to the integral since the support of the target density (numerator) is included in the support of the importance function $p_{\mathrm{IS},t}$ (Robert and Casella, 1999). Moreover, the choice of $p_{\mathrm{IS},t}$ that minimises the estimates' variance is the target density itself, and importance functions having thinner tails than the target produce estimates with infinite variance. The need is thus to find a $p_{\mathrm{IS},t}$ as close as possible to each target (per particle m) but whose tail is heavier than every one. We propose to use

$$p_{\mathrm{IS},t}(X_t, V_t, \theta_1, \theta_2) = p(y_t|X)\, p(X|V, \theta_2)\, p(\theta_1^r, \theta_2^r|\mathbf{y}, \mathbf{npp})\, f_{\mathrm{IS},t}(V_t) \tag{4.17}$$

with all the structure, except for V, given by the model (Equation 4.5). The problem of choice is thus reported to the lower-dimensional distribution $f_{\mathrm{IS},t}(V_t)$. After several tests, we use independent distributions among sites and taxa, whose shape is driven by pollen data,

$$f_{\mathrm{IS},t}(V^j|y^j_{t_0..t_n}) = \begin{cases} \mathcal{N}_0(0, \sigma^j_{\mathrm{IS}}) & \text{if} \quad y_t^j > 0 \\ 0.5\delta_0 + 0.5\mathcal{N}_0(0, \sigma^j_{\mathrm{IS}}) & \text{if} \quad y_t^j = 0 \text{ and } \exists\, t_k \quad y^j_{t_k} > 0 \\ \delta_0 & \text{if} \quad \forall\, t_k \quad y^j_{t_k} = 0 \end{cases}$$

where $N_0(0, \sigma^j_{\mathrm{IS}})$ is a Gaussian distribution truncated at 0, centred at 0 and with variance equal to σ^j_{IS}. This variance is selected as the variance of the modern NPP (**npp**) that are positive.

Random samples following the importance density Equation 4.17 are generated easily by using pieces of the MCMC sampler used for the calibration. Compared to the calibration, here, for each time t, a single X_t and V_t have to be integrated out and the values of the calibrated parameters θ_1^r and θ_2^r are selected in turn from the posterior simulations obtained in calibration.

Inference of climate over time and its parameters

We model climate as field with Markovian property in time, i.e. climate at time t depends only on previous time, noted $t-1$. The transition kernel, $p(C_t|C_{t-1}, \theta_3)$, depends on a parameter, noted θ_3, that is - now - unknown, i.e. seen as random in a Bayesian framework. Since this parameter does not depend on time, it is said to be 'static' in the context of SMC inference. We use the method described in Storvik (2002) and Fearnhead (2002) to infer θ_3. This simplifies computing by only requiring knowledge of the previous state of the algorithm but limits modelling choices to conjugate distributions. Other methods are available to deal with non-conjugated distributions (e.g. Liu and West, 2001) but they are heuristic.

We model the climate inertia by assuming that its change between two time points depends on the length of time between them. We choose, in this first approach, a Gaussian random walk model

$$p(C_t|C_{t-1}, \theta_3) = \mathcal{N}(C_{t-1}, d(t-1,t)\theta_3\Sigma_C) \tag{4.18}$$

where $d(t-1,t)$ is the length of time between t and $t-1$ and Σ_C a matrix representing the correlations between the climate variables included in C.

Following the model structure (DAG, Figure 4.2) we rewrite Equation 4.10 considering θ_3 as random,

$$\begin{aligned}
p(\theta_3, \mathrm{NPP}_{t_0:t}, C_{t_0:t}|Y_{t_0:t}) &\propto p(\theta_3, \mathrm{NPP}_{t_0:t}, C_{t_0:t}, Y_t|Y_{t_0:t-1}) \\
&= p(\theta_3, \mathrm{NPP}_{t_0:t-1}, C_{t_0:t-1}|Y_{t_0:t-1}) \\
&\qquad p(C_t|C_{t-1}, \theta_3)\, p_{\mathrm{LPJ}}(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t)\, p(Y_t|\mathrm{NPP}_t) \\
&= p(\theta_3|\mathrm{NPP}_{t_0:t-1}, C_{t_0:t-1}, Y_{t_0:t-1})\, p(\mathrm{NPP}_{t_0:t-1}, C_{t_0:t-1}|Y_{t_0:t-1}) \\
&\qquad p(C_t|C_{t-1}, \theta_3)\, p_{\mathrm{LPJ}}(\mathrm{NPP}_t|\mathrm{NPP}_{t-1}, C_t)\, p(Y_t|\mathrm{NPP}_t)
\end{aligned}$$

When the prior for θ_3 ($p(\theta_3)$, Equation 4.8) is chosen conjugated to $p(C_t|C_{t-1}, \theta_3)$, the first term of the third line is available analytically. This allows to directly obtain

the smoothed posterior distribution of θ_3 as the last filtering distribution $p(\theta_3|\text{NPP}_{t_0:t_n}, C_{t_0:t_n}, Y_{t_0:t_n})$. Moreover it only requires to save the variables from the last SMC step instead of the full particle set $\{\text{npp}^m_{t_0:t_n}, c^m_{t_0:t_n}\}$. We chose an inverse gamma prior for θ_3 which is conjugated to the normal distribution. The distribution $p(\theta_3|\text{NPP}_{t_0:t_n}, C_{t_0:t_n}, Y_{t_0:t_n}) = p(\theta_3|r_t, s_t)$ is then inverse gamma with parameters r_t and s_t updated sequentially.

The smoothed posterior distribution for θ_3 is obtained at the last iteration (time t_n). This implies that filtering distributions for climate and vegetation depend on an evolving (filtering) distribution for θ_3. A cheap improvement of the reconstruction that fits in the two step process presented previously consists in re-running the particle filter for the whole core and fixing the distribution of θ_3 to $p(\theta_3|r_{t_n}, s_{t_n})$.

Reconstruction from several cores at the same time

In this model and inference method, the reconstruction of past climate from several cores at the same time is theoretically straightforward. Suppose we have l spatially close cores sampled at sites $s_1, .., s_l$. Each one is dated and the concatenation of all the dates (potentially not aligned) forms the chronology $t_0 : t_n$. Following the DAG in Figure 4.3 and the same reasoning as for a single core, the multiple core reconstruction is obtained as

$$\begin{aligned} &p(\text{NPP}_{t_0:t_n,s_1:s_l}, C_{t_0:t_n,s_1:s_l}, \theta_3|y_{t_0:t_n,s_1:s_l}) \\ &\quad\propto p(C_{t_0:t_n,s_1:s_l}|\theta_3)\, p(\theta_3) \left(\prod_{s=1}^{l} p_{\text{LPJ}}(\text{NPP}_{t_0:t_n,s}|C_{t_0:t_n,s})\right) \prod_{t=t_0}^{t_n} p(y_{t,s(t)}|\text{NPP}_{t,s_1:s_l}) \end{aligned} \tag{4.19}$$

The differences between this equation and Equation 4.8 are, (first term) the spatio-temporal model needed for climate, (third term) the product over the vegetation dynamics for each site and (last term) the potentially multiple climate data which, now

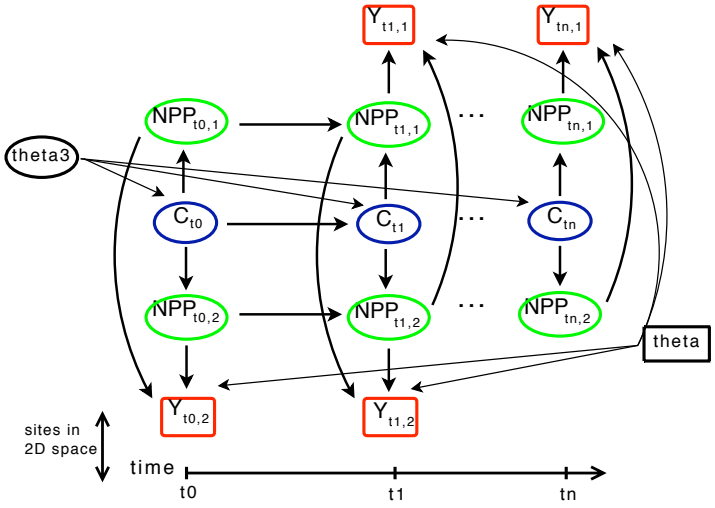

Figure 4.3: Directed Acyclic Graph of the model used for multiple core reconstruction. Climate C_t is a spatio-temporal process indexed over time. Increasing time t_0 to t_n are the ages of the core points. Two sites are represented above and below climate. For certain time points, pollen (Y) is sampled in only one site (t_0 and t_n) for others it is sampled in both. Parameters $\theta = (\theta_1, \theta_2)$ is known from the calibration, i.e. distributed following $p(\theta|\mathbf{y}, \mathbf{npp})$. Parameter θ_3 drives the climate spatio-temporal process and is to be reconstructed.

depend on the vegetation dynamics at all the sites. This term can be expanded following

$$\begin{aligned} p(Y_{t,s(t)}|\mathrm{NPP}_{t,s_1:s_l}) &= \int p(Y_{t,s(t)}|\mathrm{NPP}_{t,s_1:s_l}, \theta)\, p(\theta|\mathbf{y}, \mathbf{npp})\, d\theta \\ &= \int \left(\prod_{s\in s(t)} p(Y_{t,s}|X_{t,s})\, p(X_{t,s}|V_{t,s_1:s_l}, \theta_2) \right) \left(\prod_{s=s_1}^{s_l} p(V_{t,s}|\mathrm{NPP}_{t,s}, \theta_1) \right) \\ &\quad \left(\prod_{s=s_1}^{s_l} d\mathrm{NPP}_{t,s} \right) \left(\prod_{s\in s(t)} dX_{t,s} \right)\, p(\theta_1, \theta_2|\mathbf{y}, \mathbf{npp})\, d\theta_1\, d\theta_2 \end{aligned} \tag{4.20}$$

Thus, the integration is now performed over l variables V, and a number equal to $\dim(s(t))$ (the number of pollen points at t) of variables X. This increase in dimension makes the MCMC algorithm for the 'omnibus' weighting slower to converge (not critical

because it is very fast) and the IS approximation poorer. This later problem is critical and will be discussed in the application. For constructing the importance density, the same heuristic as in the single core case is used and the whole set of cores is considered, i.e. the function is of the form $f_{\mathrm{IS},t,s}(V_s^j|y_{t_0:t_n,s_1:s_l}^j)$.

4.4 Application: Holocene climate in South Sweden

We reconstruct three climate variables, namely January and July temperatures ($\mathrm{T_{jan}}$, $\mathrm{T_{jul}}$ in °C) and annual precipitation ($\mathrm{P_{ann}}$ in mm) in southern Sweden and over the Holocene, from pollen assemblages sampled in four close cores.

The maximum distance between cores is 400km and their time coverage is approximately the Holocene (Figure 4.4). They have been selected in the European pollen database (`www.europeanpollendatabase.net`) for their high sampling resolution, long time coverage and proximity that should allow a spatio-temporal reconstruction, i.e. a climate reconstruction from all the cores at the same time. These cores have been collected in Lake Bjärsjöholmssjön (called Mabo Moss herefater, Göransson, 1991), Lake Trummen (Trummen, Digerfeldt, 1972), Lake Flarken (Flarken, Digerfeldt, 1977) and Lake Ljustjärnen (called Gloppsjon herefater, Almquist-Jacobson, 1994). The pollen diagrams are presented in Appendix C.

The calibration of (θ_1, θ_2) is realised using an European modern pollen dataset. For specific information about the datasets, LPJ-GUESS parameters and climate interpolation, see Miller et al. (2008) and the previous chapters of this manuscript. Two models, having Poisson and negative binomial accumulation (Equations 4.6 or 4.7) are fitted. The posterior simulations required several days of computing and are stored for propagating the calibration uncertainties to reconstruction.

The SMC algorithm used for reconstruction is composed of two passes using 1000 particles. Each pass required only between 15 and 32 hours of computing (depending

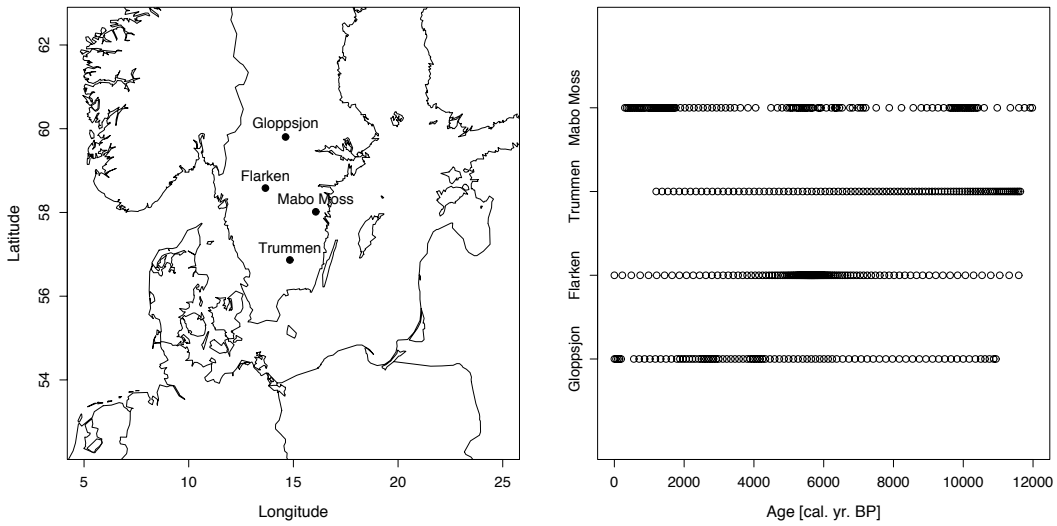

Figure 4.4: Spatial repartition and temporal coverage of the four lacustrine cores in southern Sweden. (right) The dates of each core points are plotted against the sites name.

on the machine), thanks to the parallelising of LPJ-GUESS code. We use the same proposal for all cores, obtained using the Modern Analogue Technique (MAT, Hutson, 1980; Overpeck et al., 1985; Guiot et al., 1993; Jackson and Williams, 2004). Its mean is the local smoothing of the three climatic variables (T_{jan}, T_{jul} and P_{ann}) reconstructed for each core point using the MAT. The proposal standard deviations for each variable are $(5^oC, 5^oC, 60\%)$, i.e. very large to allow an extensive climate scanning. The number of modern analogues is at maxima 7 and the distance for selection is computed using cross-validation (Guiot, 1990; Williams and Shuman, 2008). For the second pass, we fix the $p(\theta_3)$ distribution to $p(\theta_3|Y_{t_0:t_n})$. The prior mean and variance for climate are the mean and 2 times the variance of the posterior distributions obtained at the first pass. The numerical integrations required for the particle weighting (section 4.3.1) are computed offline, i.e. before the SMC algorithm run, and the $M = 1000$ 'omnibus' MCMC samples (V_t^r), $r = 1..M$ are stored for each core point. Instead of checking convergence of the numerous MCMC algorithms (one per core point), we prevent divergence by using a very large burn-in period of 1M (million) iterations and store one sample every 1 thousandth iteration to avoid correlation in the chain.

In the following sections we present several reconstructions obtained with different models for the pollen accumulation at a single site (Mabo Moss, section 4.4.1) and different sites with the same accumulation model (Negative binomial model, section 4.4.2).

4.4.1 Reconstructions for Mabo Moss with different accumulation models

We reconstruct climate at Mabo Moss with three different models for the vegetation-pollen link to understand the reconstruction sensitivity to (i) the calibration uncertainties $p(\theta|\mathbf{y}, \mathbf{npp})$ and (ii) the accumulation model selected: Poisson or negative binomial. The first model, called 'Poisson fixed', has the Poisson model for accumulation and $\theta = (\theta_1, \theta_2)$ fixed at a value $\hat{\theta}$ selected in the posterior simulations from calibration. In other words, the integral Equation 4.9 is not performed over θ for the particle weighting. The second model, called 'Poisson posterior', has the Poisson model for accumulation and considers the calibration posterior, i.e. $\theta \sim p(\theta|\mathbf{y}, \mathbf{npp})$. The last model, called 'NB posterior', has the negative binomial distribution for accumulation and considers the calibration posterior.

In Figure 4.5 we present the reconstruction of the three climate variables obtained after the first pass of the SMC algorithm at Mabo Moss and with the 'Poisson fixed' model. During this first pass reconstruction, the constraint on parameter θ_3 driving the climate inertia increases with the core time, i.e. it is better constrained at 0 than at 12 cal. kyr. BP (thousands of calendar years before 1950). This may partly explain the larger posterior intervals for older ages. Precipitation reconstructions are poor. This is likely to be due to the vegetation model which is known to be poorly constrained by precipitation (see discussion in the first chapter) and amplified by the fact that Swedish vegetation is controlled less by precipitation than temperature. The main

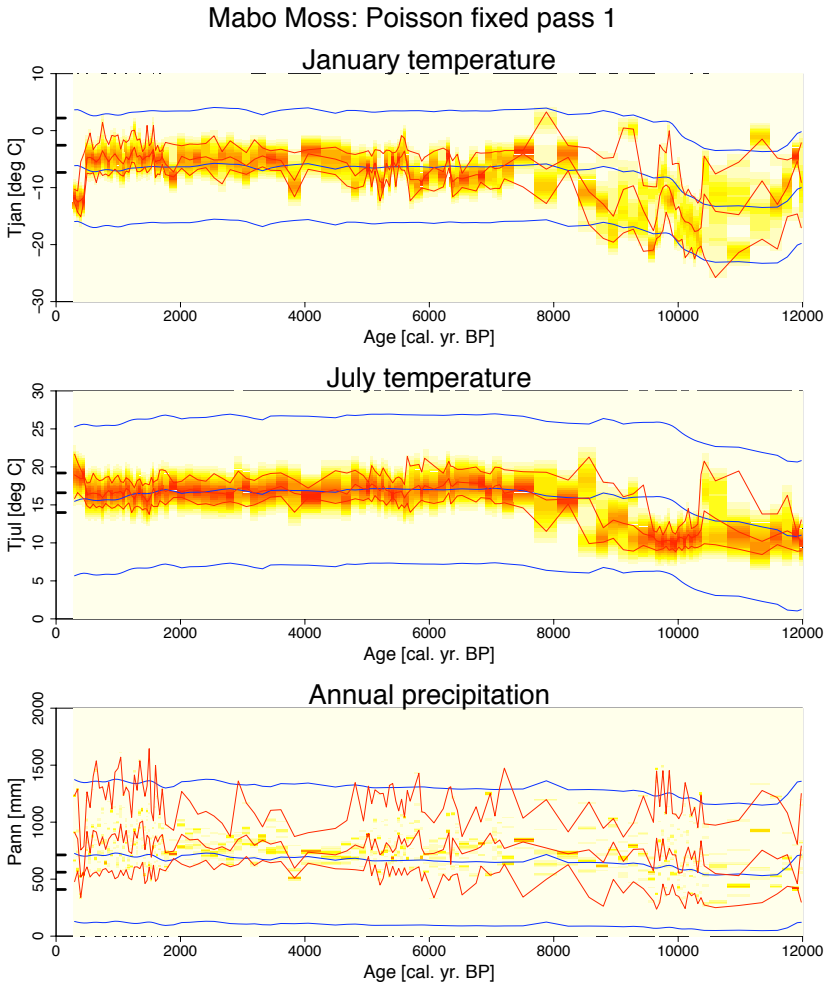

Figure 4.5: The first pass of the algorithm for climate reconstruction at Mabo Moss with the Poisson accumulation model and a fixed θ_3. (x-axis) shows the age, between 0 and 12000 cal. yr. BP (calendar years before 1950). The blue lines show the prior (obtained by MAT, see text). The colours show the posterior density (from light yellow – low densities to dark red – high densities). The red lines show the posterior 0.025, median and 0.975 quantiles. The three black marks on the left of the figures show the means and two times the standard deviation of the monthly values of the variable between the years 1901-1950

features of these reconstructions are (i) the late increase in temperature just before 8 cal. kyr. BP, which corresponds to the arrival of Fraxinus and the increase in deciduous

Quercus and Tilia (see pollen diagrams in the appendix C) and (ii) the final, strong, decrease in January temperature corresponding to the decrease or disappearing of most of deciduous tree taxa (Ulmus, Tilia, Quercus deciduous, Corylus, Betula) replaced by conifers such as Pinus and Picea and, by Grasses and Shrubs taxa (GrSh in the pollen diagram).

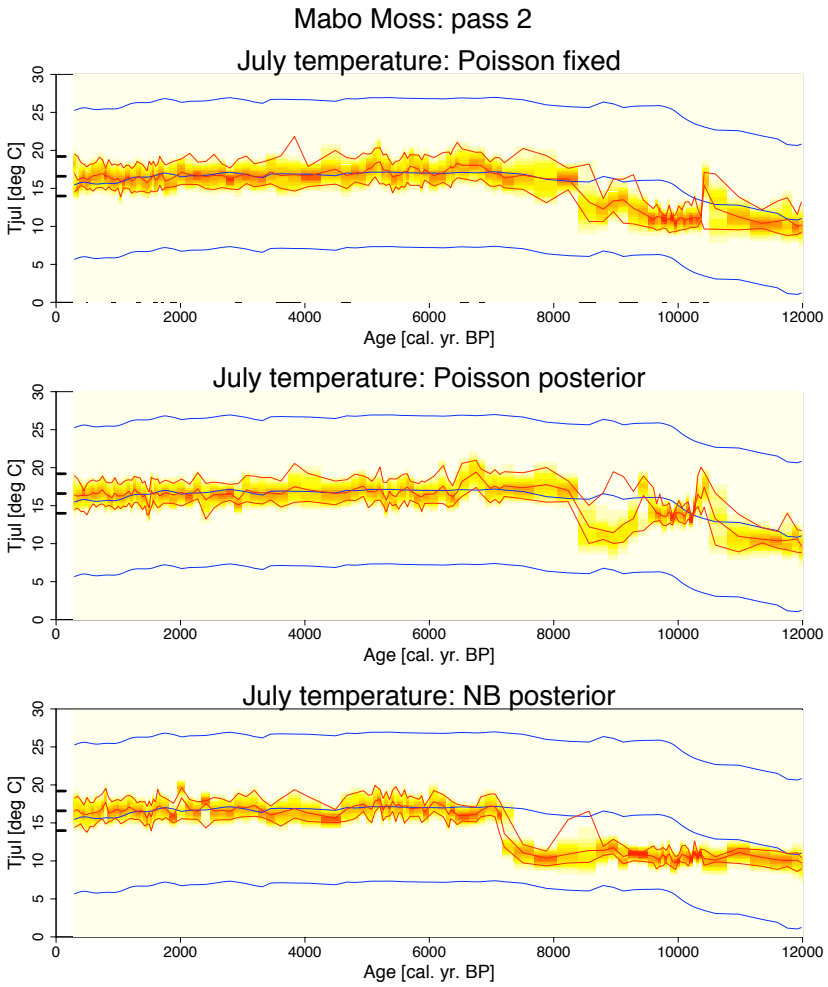

Figure 4.6: July temperature reconstructions at Mabo Moss (second pass of the algorithm) with three different models for the accumulation. The legend is the same as in Figure 4.5.

In Figure 4.6 we present the July temperature at Mabo Moss (second pass) reconstructed using the three mentioned accumulation models (Poisson with θ_3 fixed, Poisson

and negative binomial). By comparing the 'Poisson fixed' reconstruction with the one obtained previously (Figure 4.5), we see that the constraint on the θ_3 at the second pass allows to significantly reduce the uncertainties for the oldest period. The 'Poisson posterior' should theoretically include the 'Poisson fixed' and have larger variances. This is not the case, they show slightly different patterns with overlapping parts. This means that integrating in the calibration uncertainties has little influence on the posteriors compared to the errors arising from Monte Carlo integration. The 'NB' model provides a different reconstruction with a later (7.5 kyr. instead of 9 kyr. BP) temperature increase than in the Poisson models. This increase corresponds to weak changes in the pollen composition, i.e. the Quercus deciduous and Fraxinus late increase. The model seems very sensitive to very small changes in the pollen composition.

4.4.2 Reconstructions at different sites

In Figure 4.7 we present the July temperature reconstructions at the four sites using the negative binomial model. We selected this model because, even though in the Mabo Moss reconstructions it provided a reconstruction apparently less supported by data, the model validation on modern dataset (explained in the previous chapter) showed that it is better supported by modern data than the Poisson model.

Reconstructions show similar features for Mabo Moss and Flarken from one side and Trummen and Gloppsjon on the other. Mabo Moss and Flarken are located around the same latitude (see Figure 4.4). Their pollen diagrams present similar shapes that could explain the July temperature increase: starting only around 7.5 and 7 kyr. BP and related to the arrival of Fraxinus, Quercus deciduous and Tilia taxa. Gloppsjon and Trummen are located respectively north and south of these sites despite having very similar shapes in their pollen diagrams and climate reconstructions. This confirms the high sensitivity of this model to weak changes in the pollen composition. Part of this over-sensitivity may be explained by a poorer integration by the 'omnibus' sample compared to the Poisson model. But since reconstructions are coherent between cores showing the same vegetation dynamics, this feature is not only due to 'noise' from Monte Carlo integration.

Spatio-temporal reconstruction

We tried to reconstruct a global climate for these four sites in a single reconstruction exercise, i.e. a spatio-temporal reconstruction as presented section 4.3.1. We then selected the same climate model as for a single site (proximity of the sites supports this idea). The reconstruction (not shown here) crashed because the vegetations in the various pollen diagrams are significantly different. For example, the arrival and increase of deciduous taxa is significantly different between sites around 8 kyr BP. Therefore, for certain points (particularly around 8 kyr. BP), the algorithm did not manage to find a trade-off in climate for explaining the various vegetation. The integrations by IS for particle weighting produced a number of equally weighted particles having very differ-

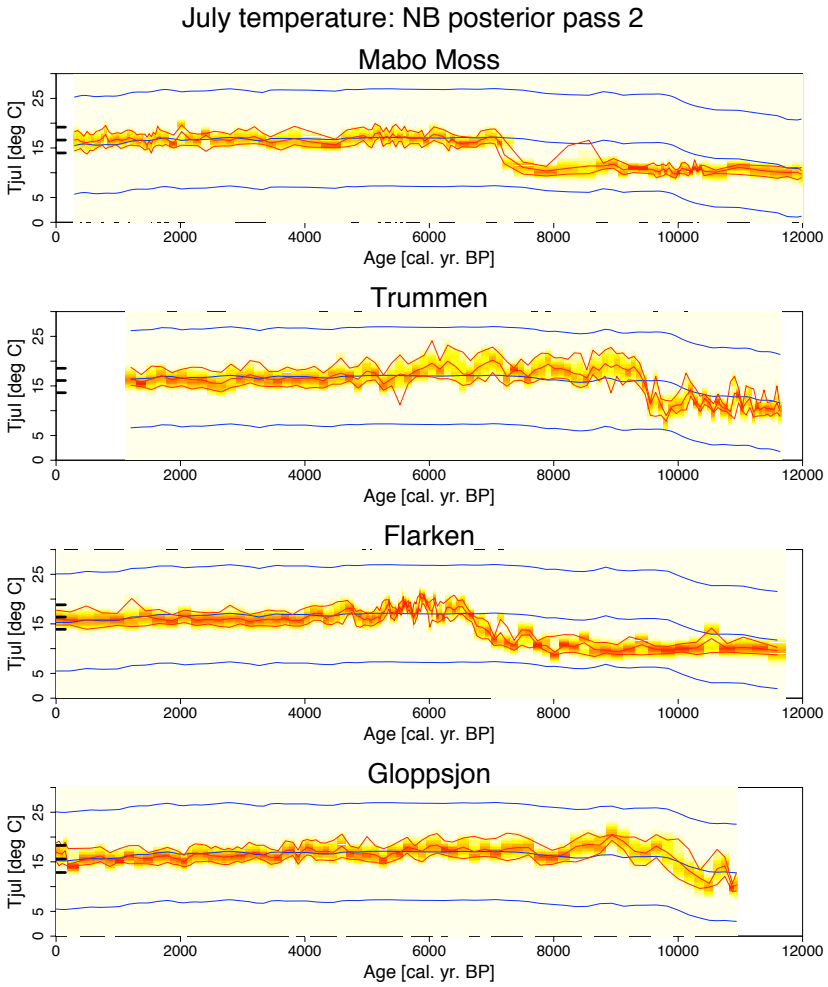

Figure 4.7: July temperature reconstructions at the four sites (second pass of the algorithm) with the negative binomial model for the accumulation. The legend is detailed in Figure 4.5.

ent and often implausible climates that indicate the failure of the algorithm. We tried to artificially reduce the prior range for climate but the posterior particle distribution always indicated a lack of convergence.

This is coherent with the pollen sequences and the reconstruction previously produced. From our model point of view, the pollen sequences of Mabo Moss and Flarken on one side and Trummen and Gloppsjon on the other cannot be the result of the same past climate dynamics – yet they must be in reality, due to their closeness in space.

This imply one must consider other processes, not included in our statistical model, nor in DVM such as LPJ-GUESS, particularly those attached to the vegetation spatial dynamics (e.g. migration).

4.5 Discussion

We proposed a semi-mechanistic model based on the representation of the processes linking climate and pollen sampled in sediments. In its higher level, it integrates a DVM for linking climate and vegetation. The following levels are modelled stochastically and represent the anthropogenic disturbance or the DVM error, the processes of pollen production and dispersion and finally the accumulation and sampling in the natural trap. We used the most up-to-date Bayesian tools for the inference of this model's parameters and past climate for four sediment cores. This TF, intrinsically based on the validity of the vegetation model and the 'completeness' of the statistical model representing the vegetation-pollen relation appeared to suffer from *structural* errors and problems with its *inference*. We define structural errors as gaps in the model structure that cause it to produce climate reconstructions that are too sensitive to changes in pollen composition, e.g. too different between sites that have certainly experienced the very similar past climate conditions. The inference limitations in calibration and reconstruction are numerous and mainly (i) preclude the proper calibration and testing of the TF and (ii) they induce a high level of noise in the reconstructions.

The structural errors may arise from inadequacy of the statistical model with respect to representing the uncertainties in the relationship between vegetation and pollen. The validation of the Poisson and Negative binomial models are one aspect of the problem but uncertainties remain on the hypotheses we made about constancy in space of the pollen production, dispersal and vegetation error. On its side, LPJ-GUESS contains a huge number of parameters that are fixed and can also contain errors. Especially, the parameters controlling the vegetation dynamics are very poorly constrained due to the lack of large temporal records of climate and vegetation that would allow their calibration. Both types of structural errors require further - deep - consideration and testing that are currently very restricted due to problems in inference.

The inference of such a semi-mechanistic TF is very limited and requires extra work in statistics. The structural problems require further investigations of the model adequation in its various levels. The vegetation model calibration may be redone or im-

proved using the modern pollen and climate dataset and the different hierarchical levels should be tested using robust procedures such as cross-validation. This is currently not permitted by the computational cost of such methods straightforwardly applied to semi-mechanistic and hierarchical and spatio-temporal structures. These problems require major changes in the inference and/or modelling strategy. Possible solutions include the use of 'emulators', i.e. statistical models copying the vegetation model structure.

The mix of a mechanistic model and stochastic relations allowed us to provide a model including a number of processes expected to link climate and pollen through space and time. For palaeoclimatology, this process-based structure is deeply original in that it fully integrates the up-to-date knowledge about climate-plant-pollen relations. It is not explored in this manuscript but the framework and the model have potentially wider applications than palaeoclimate reconstruction; for example, vegetation model calibration on the modern dataset and past vegetation dynamics reconstruction from pollen sequences could be envisioned after improvement of the statistical inference framework.

Acknowledgements

I thank Joël Chadœuf (BSP, INRA Avignon) for the discussion around multinomial overdispersion, Philippe Dussouillez (CEREGE) for its help in parallelising LPJ-GUESS and Simon Brewer for help with the dataset extraction. This works has been funded by the European Science Foundation (ESF) under the EUROCORES Programme EuroCLIMATE (project DECVEG), and by the French Centre de la Recherche Scientifique (CNRS).

Conclusion and perspectives

We built and used a full process-based TF for the reconstruction of palaeoclimate from pollen data. The approach is based on three major points, (i) the use of a DVM to link the plant species environment to their productivity, which requires the DVM inversion for palaeoclimate reconstruction, (ii) the building of a statistical hierarchical model for describing the processes linking vegetation and pollen and, (iii) the development and use of a Bayesian statistical framework to properly define the inference and include uncertainties attached to the calibration/reconstruction process in palaeoclimatology.

• The modelling of the plant-environment interactions has been achieved by plugging an up-to-date DVM, LPJ-GUESS, inside the TF, which is an improvement of what Guiot et al. (2000) have done with a static vegetation model. With such mechanistic model in the TF, palaeoclimate reconstructions require the 'dynamic' inversion of the DVM, i.e. a joint (opposed to independent) climate reconstruction based on all pollen samples from a given sediment core. In statistics, the problem translates into the inference of hidden states from a state-space model, whose core structure is implicit (only available through simulations). No classical solution exists, and we proposed to adapt a sequential Monte Carlo algorithm to obtain the reconstruction as the filtering posterior distribution. This result is a first step toward obtaining the smoothed posterior, i.e. reconstructions including the whole temporal information.

• The plant-pollen link is a statistical model representing major processes through a chain of levels. The higher level, immediately linked to the DVM outputs, represents the DVM 'errors' compared to actual vegetation, e.g. those created by local and anthropogenic disturbances. The second level models the processes of pollen production,

dispersal and accumulation in the lake. The last level represents the processes of sampling, recognising and counting a fraction of the pollen grains present in the sediment. The whole model is comparable to those used in palaeoecology for reconstructing past vegetation composition from pollen samples. However, here, it is framed in the Bayesian paradigm, allowing its inference in a single step, and it is adapted to the continental scale. From a statistical point of view, the novelty consists in the modelling of multinomial overdispersion in the context of many zeros. This huge spatial model, applied to the European dataset, raised inference and testing problems that we addressed by parallelising a MCMC algorithm and combining several model checking criteria. Despite its stochastic and process-based structure, it is not fully adequate to represent the variability present in the modern dataset. Fast and robust validation methods still need further developments to locate failures in modelling. The re-calibration of the vegetation model may form part of the solution.

- The whole process-based modelling has been framed in the Bayesian paradigm. This inference framework provides a theoretical sound basis for reconstructing past climate from TF that describe - causatively - pollen as a function of past environment. The Bayesian link between 'calibration' and 'reconstruction' processes through the so-called prior/posterior rationale allows to propagate the uncertainties between these two steps. It also opens the door to one of the most active fields of research in modern inference techniques around Monte Carlo methods.

The process-based approach is aimed at complementing the correlative TF by being based on independent hypotheses and integrating up-to-date knowledge in vegetation modelling. Its purpose is to provide independent reconstructions, increasing the confidence in palaeoclimate reconstructions that are only available through the TF. We believe that, by enhancing the similarities between Species Distribution Models (SDM) used in Ecology and TF, the huge modelling and inference effort required to obtain improved process-based approaches could be shared between disciplines.

Perspectives

The uses of our TF in calibration and reconstruction were strongly restricted by the lack of rapid inference and validation methods. This arose from the semi-mechanistic and spatio-temporal nature of the TF. The development of inference and validation procedures for such approach forms a huge field of research having a wide range of applications outside palaeoclimatology and ecology.

As discussed in the calibration results, the DVM lacks quantitative calibration and testing based on the modern species distributions. The statistical model we used for the calibration of the vegetation-pollen relation could be a support for such comparisons. When the inference of semi-mechanistic models will be partly solved, the approach we presented would allow (a) the calibration of parameters inside the DVM, based on modern species distributions, (b) the modelling of anthropogenic disturbances and spatial vegetation dynamics, and in 'reconstruction', (c) the joint calibration and reconstruction of spatio-temporal vegetation dynamics on large spatio-temporal scales.

Inference of semi-mechanistic models

Throughout the whole work we faced problems with the inference and the validation. They are due to the mechanistic DVM defining an implicit distribution and, the use of spatio-temporal structures for large datasets. For example, on the DVM side, past climate reconstructions have been only obtained as filtering distributions, cross-validation of the reconstruction method was only possible on a very small set of sites and the DVM parameters calibration was ignored due to the too highly dimensional integration required. On the statistical side, the validation of the spatial vegetation-pollen relation has been considerably constrained by the computing time required for the inference.

The inference of spatio-temporal structures on very large datasets is already the subject of many statistical researches in modelling and inference (e.g. in spatial statistics, Calder and Cressie, 2007; Fuentes, 2007; Banerjee et al., 2008; Zhang and Wang, 2009). We believe that these approaches may be integrated in the future to allow faster

inference in our approach.

The inference and validation for semi-mechanistic models are in their infancy and require deep consideration in statistics, because they have a tremendous range of applications. Indeed, from genetics and molecular biology to vegetation and climate modelling, there is a need for calibrating parameters inside mechanistic models or reconstructing states of hidden variables linked through mechanistic models from diverse sources of information.

Recent propositions for tackling the problem have been made. For example, in genetics, ABC methods (Beaumont et al., 2002; Sisson et al., 2006; Beaumont et al., 2010, and references therein) allow to calibrate a few parameters inside simulatory models based on highly dimensional genetic data. These methods have recently been applied to the inference of parameters in a biological dynamic system driven by a few differential equations (Toni et al., 2009). Although this is similar to our problem, this remains far from reconstructing past states of the system when the mechanistic model is as computationally demanding as ours. In Global Circulation Models (GCM), earth system models and ecosystems models, the techniques related to 'data-assimilation' or 'data-model fusion' deal with the integration of models and data for the calibration of models or the correction of their predictions (e.g. Daley, 1991; Hargreaves and Annan, 2002; Raupach et al., 2005). In these cases, the models are often over computationally demanding (e.g. GCM) and statistical modelling is reduced to very simple Gaussian models, either because this is the only way for obtaining a tractable inference or because the data are supposed to be 'close' to quantities simulated from these models (compared to the pollen as an indicator of the vegetation). Recent propositions such as model emulation and 'reified' modelling (Currin et al., 1991; Kennedy and O'Hagan, 2001; Goldstein and Rougier, 2009) provide new perspectives for the modelling and inference of such highly complicated physical systems by emulating them using computationally inferable statistical models. The weakness of this approach is its use of a statistical model, the emulator, which evolves in its own statistical world, restricting feedbacks and improvements to the mechanistic models in use (DVM in our case).

We believe that by combining several of these approaches, in a near future, solutions

will be available for continuing the development and improvement of our process-based TF.

Validation and improvement of vegetation models

Validation is the Achille's heel of our semi-mechanistic approach. In this work oriented on palaeoclimatology, we considered the vegetation model as valid (i.e. properly calibrated and tested) for the modelling of modern vegetation distributions although (i) it has not been validated on dataset that are as extensive and quantitatively precise as our European modern pollen dataset and (ii) it does not account for spatial vegetation dynamics, such as migration and plant dispersion processes.

When the inference of semi-mechanistic models will be partly solved, the framework we presented in the last chapter would readily allow the calibration of parameters inside the DVM and its proper testing as a component of the hierarchical statistical model. Such inference has to be discussed and several parameters must remain fixed since they are derived from theory and would lose their meaning in a 'blind' statistical fitting. Moreover, the DVM does not model any spatial component (e.g. species migrations) nor anthropogenic disturbances, which make it inappropriate for the direct modelling of the modern species distributions that are expected to feature footprints from these processes. In the statistical model linking the DVM outputs to pollen data, we crudely took part of these processes into account by assuming independent errors (in space and between species) between the simulated and the actual vegetation featured by pollen data. The imperfect fit of the statistical model to pollen data is due to this crude representation of the vegetation spatial processes and, relatedly to the DVM.

The improved inference framework for the semi-mechanistic model we proposed would then help locating gaps in the DVM to represent the modern species (quantitative) distribution. It would also allow the calibration of future spatial components, forming the base for the next generation of vegetation models. This migration component cannot be fully fitted on modern data since they cover a very restricted temporal range. The pollen sequences collected in hundreds of cores around the world provide

a unique information about these dynamics. They could be used to calibrate the vegetation dynamics at the same time as climate reconstruction are performed based on many cores.

To conclude, palaeoclimatology and palaeoecology bear the keys for the next generation of spatio-temporal dynamical vegetation models, which are the most credible tools for the prediction of future vegetation under climate having modern no analogues.

Bibliography

Aitchison, J. (1982). The statitical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2):139–177.

Alley, R. B. (2001). The key to the past? *Nature*, 409(289):doi:10.1038/35053245.

Almquist-Jacobson, H. (1994). Interaction of the Holocene climate, water balance, vegetation, fire, and cultural land-use in Swedish Borderland. *Lundqua Thesis*, 30:1–82.

Ammann, B., Birks, H. J. B., Brooks, S. J., Eicher, U., von Grafenstein, U., Hofmann, W., Lemdahl, G., Schwander, J., Tobolski, K., and Wick, L. (2000). Quantification of biotic responses to rapid climatic changes around the Younger Dryas — a synthesis. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 159.

Ancelet, S., Etienne, M.-P., Benoît, H., and Parent, E. (2009). Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, online.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 72(2):1–33.

Atkinson, T. C., Briffa, K. R., Coope, G. R., Joachim, M. J., and Perry, D. (1986). Climatic calibration of coleopteran data. In Berglund, B. E., editor, *Handbook of Holocene Palaeoecology and Palaeyhydrology*, pages 851–858. Wiley and Sons, Chichester.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *J. Roy. Statist. Soc. Ser. B*, 70:825–848.

Barry, R. P. and Ver Hoef, J. M. (1996). Blackbox Kriging: Kriging without specifying variogram models. *Journal of Agricultural, Biological and Environmental Statistics*, 1(3):297–322.

Bartlein, P. J., Prentice, I. C., and Webb III, T. (1986). Climatic response surfaces from pollen data for some eastern north american taxa. *Journal of Biogeography*, 13(35-57).

Bartlein, P. J., Webb III, T., and Fleri, E. (1984). Holocene climatic change in the northern Midwest: Pollen-derived estimates. *Quaternary Research*, 22:361–374.

Beaumont, M., Robert, C. P., Marin, J.-M., and Cornuet, J. M. (2010). Adaptivity for ABC algorithms: the ABC-PMC scheme. *Biometrika*, (to appear).

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162:2025–2035.

Berger, A. and Loutre, M. F. (2004). Théorie astronomique des paléoclimats. *C. R. Geoscience*, 336:701–709.

Berger, A. L. (1977). Support for the astronomical theory of climatic change. *Nature*, 269:44–45.

Bhattacharya, S. (2006). A bayesian semiparametric model for organism based environmental reconstruction. *Environmetrics*, 17:763–776.

Bhattacharya, S. and Haslett, J. (2004). Importance Re-Sampling MCMC for Cross-Validation in Inverse Problems. *Bayesian Analysis*, 1.

Bordon, A. (2008). *Dynamique de la végétation et variations climatiques dans les Balkans au cours du dernier cycle climatique à partir des séquences polliniques des lacs Maliq et Ochrid (Albanie).* PhD thesis, Université de Franche-Comté.

Bottema, S. (1974). *Late Quaternary Vegetation History of Northwestern Greece.* PhD thesis, Rijksuniv. Groningen.

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, J.-Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, T., Hewitt, C. D., Kageyama, M., Kitoh, A., Laîné, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S. L., Yu, Y., and Zhao, Y. (2007). Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum - Part 1: experiments and large-scale features. *Climate of the Past*.

Brewer, S., Guiot, J., and Barboni, D. (2007). Pollen data as climate proxies. In Elias, S. A., editor, *Encyclopedia of Quaternary Science*, volume 4, pages 2498–2510. Elsevier.

Broström, A., Nielsen, A. B., Gaillard, M. J., Hjelle, K., Mazier, F., Binney, H., Bunting, J., Fyfe, R., Meltsov, V., Poska, A., Räsänen, S., Soepboer, W., von Stedingk, H., Suutari, H., and Sugita, S. (2008). Pollen productivity estimates of key european plant taxa for quantitative reconstruction of past vegetation: a review. *Vegetation History and Archæobotany*, DOI 10.1007/s00334-008-0148-8.

Brugiapaglia, E. (1996). *Dynamique de la végétation tardiglaciaire et holocène dans les Alpes italiennes nord-occidentales.* PhD thesis, Université Aix-Marseille III.

Bugmann, H. (2001). A review of forest gap models. *Climatic Change*, 51:259–305.

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validation method for dependent data. *Biometrika*, 81:351–358.

Calder, C. A. and Cressie, N. (2007). Some topics in convolution-based spatial modelling. In *Proceedings of the 56th Session of the International Statistics Institute*.

Cappé, O., Guillin, A., Marin, J. M., and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Model*. Springer Texts in Statistics. Springer, New York.

Chase, J. M. and Leibold, M. A. (2003). *Ecological Niches: Linking Classical and Contemporary Approaches*. Interspecific Interactions. The University of Chicago Press, Chicago and London.

COHMAP Members (1988). Climatic changes of the last 18 000 years: Observations and model simulations. *Science*, 241:1043–1052.

Cowling, S. A. and Sykes, M. T. (1999). Physiological significance of low atmospheric co_2 for plant-climate interactions. *Quaternary Research*, 52:237–242.

Cramer, W. (2002). Biome models. In Mooney, H. A. and Canadell, J. G., editors, *Encyclopedia of Global Environmental Change*, volume 2, The Earth system: biological and ecological dimensions of global environmental change. John Wiley and Sons.

Cramer, W., Bondeau, A., Woodward, F. I., Prentice, I. C., Betts, R. A., Brovkin, V., Cox, P. M., Fisher, V., Foley, J. A., Friend, A. D., Kucharik, C., Lomas, M. R., Ramankutty, N., Sitch, S., Smith, B., White, A., and Young-Molling, C. (2001). Global response of terrestrial ecosystem structure and function to co_2 and climate change: results from six dynamic global vegetation models. *Global Change Biology*, 7(4):357–373.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley and Sons, New York.

Cressie, N., Calder, C., Clark, J., Ver Hoef, J., and Wikle, C. (2007). Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Technical Report 798, Ohio State University.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86:953–963.

Daley, R. (1991). Atmospheric data analysis. In Dessler, A. J., Houghton, J. T., and Rycroft, M. J., editors, *Cambridge Atmospheric and Space Science Series*. Cambridge University Press, Cambridge.

Davis, M. B. (1963). On the theory of pollen analysis. *American Journal of Science*, 261:897–912.

Digerfeldt, G. (1972). The post-glacial development of lake Trummen. Regional vegetation history, water level and palaeolimnology. *Folia Limnologia Scandinavica*, 16:1–96.

Digerfeldt, G. (1977). The flandrian development of lake Flarken. Regional vegetation history and palaeolimnology. *University of Lund Department of Quaternary Geology. Report*, 13:1–101.

Diggle, P. J., Ribeiro Jr, P. J., and Christensen, O. F. (2003). An introduction to model-based geostatistics. In Moller, J., editor, *Spatial statistics and computational methods*, Lecture notes in statistics. Springer, New York.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Statistics for Ingineering and Information Science. Springer, New York.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.*, 40:677–697.

Faegri, K. and Iversen, J. (1964). *Textbook on pollen analysis*. Hafner, New York.

Fearnhead, P. (2002). Markov chain monte carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862.

Fuentes, M. (2007). Appropriate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102:321–331.

Gaillard, M. J., Sugita, S., Bunting, J., Dearing, J., and Bittmann, F. (2008). Human impact on terrestrial ecosystems, pollen calibration and quantitative reconstruction of past land-cover. *Vegetation History and Archaeobotany*, 17:415–418.

Garreta, V., Miller, P. A., Guiot, J., Hély, C., Brewer, S., Sykes, M. T., and Litt, T. (2009). A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model. *Climate Dynamics*, doi: 10.1007/s00382-009-0629-1.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, second edition.

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assement of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gersonde, R., Crosta, X., Abelmann, A., and Armand, L. (2005). Sea-surface temperature and sea ice distribution of the Southern Ocean at the EPILOG Last Glacial Maximum. *Quaternary Science Reviews*, 24:869–896.

Gerten, D., Schabhoff, S., Haberlandt, U., Lucht, W., and Sitch, S. (2004). Terrestrial vegetation and water balance an hydrological evaluation of a dynamic global vegetation model. *Journal of Hydrology*, 286:249–270.

Goldstein, M. and Rougier, J. (2009). Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, 138:1221–1239.

Gonzales, L. M., Williams, J. W., and Grimm, E. C. (2009). Expanded response-surfaces: a new method to reconstruct paleoclimates from fossil pollen assemblages that lack modern analogues. *Quaternary Science Reviews*, 28:3315–3332.

Göransson, H. (1991). Vegetation and man around lake Bjärsjöholmssjön during prehistoric time. *Lundqua Report*, 31.

Guiot, J. (1990). Methodology of the last climatic cycle reconstruction from pollen data. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 80(1):49–69.

Guiot, J., De Beaulieu, J. L., Cheddadi, R., David, F., Ponel, P., and Reille, M. (1993). The climate in Western Europe during the Last Glacial Interglacial Cycle derived from pollen and insect remains. *Palaeogeaography, Palaeoclimatology, Palaeoecology*, 103:73–93.

Guiot, J., De Beaulieu, J. L., Pons, A., and Reille, M. (1989). A 140,000-year climatic reconstruction from two European pollen records. *Nature*, 338:309–313.

Guiot, J. and De Vernal, A. (2007). Transfer Functions: Methods for Quantitative Paleoceanography Based on Microfossils. In De Vernal, C. H.-M. . A., editor, *Developments in Marine Geology*, volume 1, chapter C. Elsevier, Amsterdam.

Guiot, J., Hély-Alleaume, C., Wu, H., and Gaucherel, C. (2008). Interactions between vegetation and climate variability: what are the lessons of models and palaeovegetation data. *C. R. Geoscience*, 340:595–601.

Guiot, J., Torre, F., Jolly, D., Peyron, O., Borreux, J. J., and Cheddadi, R. (2000). Inverse vegetation modeling by Monte Carlo sampling to reconstruct paleoclimate under changed precipitation seasonality and CO_2 conditions: application to glacial climate in Mediterranean region. *Ecological Modelling*, 1(127):119–140.

Guiot, J., Wu, H. B., Garreta, V., Hatté, C., and Magny, M. (2009). A few prospective ideas on climate reconstruction: from a statistical single proxy approach towards a multi-proxy approach. *Climate of the Past*, 5:571–583.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8:993–1009.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2):147–186.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B*, 29:83–100.

Hargreaves, J. C. and Annan, J. D. (2002). Assimilation of paleo-data in a simple earth system model. *Climate Dynamics*, 19:371–381.

Haslett, J., Whiley, M., Bhattacharya, S., Salter Townshend, M., Wilson, S., Allen, J. R. M., Huntley, B., and Mitchell, F. J. G. (2006). Bayesian Paleoclimate Reconstruction. *Journal of the Royal Statistical Society, Series A*, 169(3):395–438.

Hatté, C., Rousseau, D.-D., and Guiot, J. (2009). Climate reconstruction from pollen and δ13c records using inverse vegetation modeling – implication for past and future climates. *Climate of the Past*, 5:147–156.

Haxeltine, A. and Prentice, I. C. (1996). BIOME3: an equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability and competition among plant functional types. *Global Biogeochemical Cycles*, 10:693–709.

Hickler, T., Smith, B., Sykes, M. T., Davis, M., Sugita, S., and Walker, K. (2004). Using a generalized vegetation model to simulate vegetation dynamics in northeastern USA. *Ecology*, 85:519–530.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5:173–190.

Huntley, B. (1996). Quaternary palaeoecology and ecology. *Quaternary Science Reviews*, 15(591-606).

Hutson, W. H. (1980). The Agulhas current during the late Pleistocene: Analysis of modern faunals analogs. *Science*, 207:64–66.

Hutton, J. (1795). *Theory of the Earth*. Cadell, Junior and Davies, London.

Imbrie, J. and Kipp, N. G. (1971). A new micropaleontological method for quantitative paleoclimatology: Application to the late Pleistocene Caribbean core. In Turekian, K. K., editor, *The late Cenozoic glacial age*. Yale University Press.

Indermuhle, A., Stocker, T. F., Joos, F., Fisher, H., Smith, H. J., Wahlen, M., Deck, B., Mastroianni, D., Tshumi, J., Blunier, T., Meyer, R., and Stauffer, B. (1999). Holocene carbon-cycle dynamics based on CO2 trapped in ice at Taylor Dome, Antarctica. *Nature*, 398:121–126.

IPCC Core Writing Team (2007). Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. In Pachauri, R. and Reisinger, A., editors, *IPCC Fourth Assessment Report*, page 104. IPCC, Geneva.

Iversen, J. (1944). *Visum*, *Hedera* and *Ilex* as climatic indicators. *Geologiska Föreninger Förhandlingar*, 66:463–483.

Jackson, S. T. and Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late quaternary. *Paleobiology*, 26(4):194–220.

Jackson, S. T. and Williams, J. W. (2004). Modern analogs in quaternary paleoecology - here today, gone yesterday, gone tomorrow? *Annual Review of Earth and Planetary Sciences*, 32:495–537.

Jolly, D. and Haxeltine, A. (1997). Effect of low glacial atmospheric co_2 on tropical African montane vegetation. *Science*, 276.

Jost, A., Lunt, D., Kageyama, M., Abe-Ouchi, A., Peyron, O., Valdes, P. J., and Ramstein, G. (2005). High-resolution simulations of the last glacial maximum climate over Europe: a solution to discrepancies with continental palaeoclimatic reconstructions ? *Climate Dynamics*, 24:577–590.

Kearney, M. (2006). Habitat, environment and niche: what are we modelling? *Oikos*, 115(1):186–191.

Kearney, M. and Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, 12.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, 63:425–464.

Koca, D., Smith, B., and Sykes, M. T. (2006). Modelling regional climate change effects on Swedish ecosystems. *Climatic Change*, 78:381–406.

Kühl, N., Gebhardt, Litt, T., and Hense, A. (2002). Probability Density Function as Botanical-Climatological Transfer Functions for Climate Reconstruction. *Quaternary Research*, 58:381–392.

L'Ecuyer, P. (1999). Good parameters and implementations for combined multiple recursive random number generators. *Operations Research*, 47(1):159–164.

Legendre, P. (1993). Spatial autocorrelation–trouble or new paradigm. *Ecology*, 74:1659–1673.

Leonard, T. (1977). Bayesiam simultaneous estimation for several multinomial distributions. *Communications in Statistics - Theory and Methods*, 6(7):619 – 630.

Litt, T., Schölzel, C., Kühl, N., and Brauer, A. (2009). Holocene vegetation and climate history in the Westeifel Volcanic Field (Germany) based on annualy laminated lacustrine maar sediments. *Boreas*, 38(4):679–690.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In Doucet, A., De Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, Statistics for Ingineering and Information Science. Springer.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo Methods for Dynamic Systems. *JASA*, 93:1032–1044.

Loader, C. (1999). *Local Regression and Likelihood*. Statistics and computing. Springer-Verlag.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22:1649–1660.

Marshall, E. C. and Spiegelhalter, D. J. (2007). Identifying outliers in bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, 2(2):409–444.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecological Letters*, 8(11):1235–1246.

McGuire, A. D., Sitch, S., and Clein, J. S. (2001). Carbon balance of the terrestrial biosphere in the twentieth century: analyses of CO2, climate and land use effects with four process-based ecosystem models. *Global Biogeochemical Cycles*, 15:183–206.

Milankovitch, M. (1941). *Kanon der Erdbestrahlung und seine Anwendung auf des Eiszeitenproblem (Canon of Insolation and the Ice Age Problem, English translation by Israel Program for the US Department of Commerce and the National Science Foundation, Washington DC, 1969, and by Zavod za Udzbenike I nastavna Sredstva, in cooperation with Muzej nauke I technike Srpske akademije nauka I umetnosti, Beograd, 1998)*, volume 33 of *section of Mathematical and natural Science, Spec. Publ. 132*. Royal Serbian Sciences.

Miller, P. A., Giesecke, T., Hickler, T., Bradshaw, R. H. W., Smith, B., Seppä, H., Valdes, P. J., and Sykes, M. T. (2008). Exploring climatic and biotic controls on holocene vegetation change in Fennoscandia. *Journal of Ecology*, 96(2):247–259.

Morin, X. and Lechowicz, M. J. (2008). Contemporary perspectives on the niche that can improve models of species range shifts under climate change. *Biology Letters*, 4(5):573–576.

Morin, X. and Thuillier, W. (2009). Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. *Ecology*, 90(5):1301–1313.

Mosimann, J. E. (1963). On the compound negative multinomial distribution and correlation among inversely sampled pollen counts. *Biometrika*, 50.

Murray, N. B., Cannell, M. G. R., and Smith, I. (1989). Date of budburst of fifteen tree species in Britain following climatic warming. *Journal of Applied Ecology*, 26:693–700.

Nakagawa, T., Tarasov, P. E., Nishida, K., Gotanda, K., and Yasuda, Y. (2002). Quantitative pollen-based climate reconstruction in central Japan: application to surface and Late Quaternary spectra. *Quaternary Science Reviews*, 21:2099–2113.

Neumann, F., Schölzel, C., Litt, T., Hense, A., and Stein, M. (2007). Holocene vegetation and climate history of the northern Golan heights (Near East). *Vegetation History and Archaeobotany*.

New, M., Lister, D., Hulme, M., and Makin, I. (2002). A high-resolution data set of surface climate over global land areas. *Climate Research*, 21.

Overpeck, J. T., Webb III, T., and Prentice, I. C. (1985). Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs. *Quaternary Research*, 23:87–108.

Paciorek, C. J. and McLachlan, J. S. (2009). Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition. *Journal of the American Statistical Association*, 104:608–622.

Parsons, R. W. and Prentice, I. C. (1981). Statistical approaches to R-values and the pollen-vegetational relationship. *Review of Palaeobotany and Palynology*, 32:127–152.

Peyron, O., Guiot, J., Cheddadi, R., Tarasov, P., Reille, M., de Beaulieu, J.-L., Bottema, S., and Andrieu, V. (1998). Climatic reconstruction in Europe for 18,000 YR B.P. from pollen data. *Quaternary Research*, 49:183–196.

PMIP Participants (2000). *PMIP, Paleoclimate Modeling Intercomparison Project: proceedings of the third PMIP workshop*. WCRP-111, WMO/TD-1007, Canada, 4-8 october.

Prentice, I. C. (1985). Pollen representation, source area and basin size: toward a unified theory of pollen analysis. *Quaternary Research*, 23:76–86.

Prentice, I. C. (1986). Vegetation responses to past climatic variation. *Plant Ecology*, 67(2):1573–5052.

Prentice, I. C. (1988). Palaeoecology and plant population dynamics. *Trends in Ecology & Evolution*, 3(12):343–345.

Prentice, I. C., Bartlein, P. J., and Webb III, T. (1991). Vegetation and climate change in eastern North America since the last glacial maximum. *Ecology*, 72(6):2038–2056.

Prentice, I. C., Bondeau, A., Cramer, W., Harrison, S. P., Hickler, T., Lucht, W., Sitch, S., Smith, B., and Sykes, M. T. (2007). Dynamic Global Vegetation Modeling: Quantifying terrestrial ecosystem responses to large-scale environmental change. In Canadell, J. G., Pataki, D. E., and Pitelka, L. F., editors, *Terrestrial Ecosystems in a Changing World*, Global Change – The IGBP Series, pages 175–192. Springer, Berlin Heidelberg.

Prentice, I. C., Cramer, W., Harrison, S. P., Leemans, R., Monserud, R. A., and Solomon, A. M. (1992). A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*, 19:117–134.

Prentice, I. C. and Harrison, S. P. (2009). Ecosystem effect of co_2 concentration: evidence from past climates. *Climate of the Past*, 5:297–307.

Prentice, I. C. and Helmisaari, H. (1991). Silvics of north European trees: Compilation, comparisons and implications for forest succession modelling. *Forest Ecology and Management*, 42:79–93.

Prentice, I. C. and Parsons, R. W. (1983). Maximum likelihood linear calibration of pollen spectra in terms of forest composition. *Biometrics*, 39:1051–1057.

Raupach, M. R., Rayner, P. J., Barret, D. J., DeFries, R. S., Heimann, M., Ojima, D. S., Quegan, S., and Schmullius, C. C. (2005). Model-data synthesis in terrestrial carbon carbon observations: methods, data requirements and data uncertainties specifications. *Global Change Biology*, 11:378–397.

Ridout, M., Demétrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros. In *International Biometric Conference*, Cape Town.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Robert, C. P. (2001). *The Bayesian Choice*. Springer, 2 edition.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York.

Rousseau, D., Hatté, C., Guiot, J., Duzer, D., Schevin, P., and Kukla, G. (2006). Reconstruction of the Grande Pile Eemian using inverse modelling of biomes and δ13c. *Quaternary Science Reviews*, 25:2808–2819.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172.

Salter Townshend, M. (2009). *Fast Approximate Inverse Bayesian Inference in non-parametric Multivariate Regression with application to palaeoclimate reconstruction*. Phd thesis, University of Dublin, Trinity College.

Salter-Townshend, M. and Haslett, J. (2006). Modelling zero inflation of compositional data. In *International Workshop on Statistical Modelling*, Galway.

Sanchez Goni, M. F. and Hannon, G. (1999). High altitude vegetational pattern on the Iberian Mountain Chain (north-central Spain) during the Holocene. *The Holocene*, 9(1):39–57.

Shugart, H. H. (1984). *A Theory of Forest Dynamics. The Ecological Implications of Forest Succession Models.* Springer, New York.

Siegenthaler, U., Monnin, E., Kawamura, K., Spahni, R., Shwander, J., Stauffer, B., Stocker, T. F., Barnola, J. M., and Fisher, H. (2005). Supporting evidence from the EPICA dronning Maud Land ice core for atmospheric CO2 change during the past millenium. *Tellus*, 57B(7):51–57.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2006). Sequential Monte Carlo without likelihoods. *PNAS*.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J., Levis, S., Lucht, W., Sykes, M., Thonicke, K., and Venevsky, S. (2003). Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ Dynamic Global Vegetation Model. *Global Change Biology*, 9:161–185.

Smith, B., Prentice, I. C., and Sykes, M. T. (2001). Representation of vegetation dynamics in modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. *Global Ecology & Biogeography*, 10:621–637.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., editors (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

St. Jacques, J. M., Cumming, B. F., and Smol, J. P. (2008). A pre-European settlement pollen–climate calibration set for Minnesota, USA: developing tools for palaeoclimatic reconstructions. *Journal of Biogeography*, 35:306–324.

Stein, M. L. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, 37:3–10.

Stern, H. S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19:2377–2397.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on signal Processing*, 50(2).

Sugita, S. (2007a). Theory of quantitative reconstruction of vegetation I: pollen from large sites REVEALS regional vegetation composition. *The Holocene*, 17(2):229–241.

Sugita, S. (2007b). Theory of quantitative reconstruction of vegetation II: all you need is LOVE. *The Holocene*, 17(2):243–257.

Sykes, M. T., Prentice, I. C., and Cramer, W. (1996). A bioclimatic model for the potential distribution of northern European tree species under present and future climates. *Journal of Biogeography*, 23:203–233.

Tauber, H. (1965). Differential pollen dispersion and the interpretation of pollen diagrams. *Danmarks Geol. Undersœgelse II.*, 89.

Telford, R. J. (2006). Limitations of dinoflagellate cyst transfer functions. *Quaternary Science Reviews*, pages 1375–1382.

Telford, R. J. and Birks, H. J. B. (2005). The secret assumption of transfer functions: problem with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews*, 24:2173–2179.

Telford, R. J. and Birks, H. J. B. (2009). Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, 28:1309–1316.

ter Braak, C. J. F., Juggins, S., Birks, H. J. B., and Van der Voet, H. (1993). Weighted averaging partial least squares regression (WA-PLS): Definition and comparison with other methods for species-environment calibration. In Patil, G. P. and R., R. C., editors, *Multivariate environmental statistics*. Elsevier, Amsterdam.

Thompson, R. S., Anderson, K. H., and Bartlein, P. J. (2008). Quantitative estimation of bioclimatic parameters from presence/absence vegetation data in north america by the modern analogue technique. *Quaternary Science Reviews*, 27:1234–1254.

Thonicke, K., Venevsky, S., Sitch, S., and Cramer, W. (2001). The role of fire disturbance for global vegetation dynamics: coupling fire into a Dynamic Global Vegetation Model. *Global Ecology & Biogeography*, 6:483–495.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. Roy. Soc. Interface*, 6(31):187–202.

Vasko, K., Toivonen, H. T., and Korhola, A. (2000). A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction. *Journal of Paleolimnology*, 24(2):43–250.

von Post, L. (1916). Om skogsträdpollen i sydsvenska torfmosslagerföljder. *Geologiska föreningens Stockholm förhandlingar*, 38:384–390.

Webb, T. (1974). Corresponding distributions of modern pollen and vegetation in Lower Michigan. *Ecology*, 55(17-18).

Webb, T. J. and Bryson, R. A. (1972). Late- and Postglacial climatic change in the northern Midwest, USA: Quantitative estimates derived from fossil pollen spectra by multivariate statistical analysis. *Quaternary Research*, 2:70 – 115.

Webb III, T. (1986). Is vegetation in equilibrium with climate? how to interpret late-Quaternary pollen data. *Plant Ecology*, 67:75–91.

Whitmore, J., Gajewski, K., Sawada, M., Williams, J. W., Shuman, B., Bartlein, P. J., Minckley, T., Viau, A. E., Webb III, T., Anderson, P. M., and Brubaker, L. B. (2005). North american and greenland modern pollen data for multi-scale paleoecological and paleoclimatic applications. *Quaternary Science Reviews*, 24:1828–1848.

Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A., and Snyder, M. A. (2009). Niches, models, and climate change: Assessing the assumptions and uncertainties. *PNAS*, 106(2):19729–19736.

Williams, J. W. and Jackson, S. T. (2007). Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment*, 5(9):475–482.

Williams, J. W., Post, D. M., Cwynar, L. C., Lotter, A. F., and Levesque, A. J. (2002). Rapid vegetation responses to past climate change. *Geology*, 30:971–974.

Williams, J. W. and Shuman, B. (2008). Obtaining accurate and precise environmental reconstructions from the modern analog technique and North American surface pollen dataset. *Quaternary Science Reviews*, 27:669–687.

Williams, J. W., Shuman, B., Bartlein, P. J., Whitmore, J., Gajewski, K., Sawada, M., Minckley, T., Shafer, S., Viau, A. E., Webb III, T., Anderson, P. M., Brubaker, L. B., Whitlock, C., and Davis, O. K. (2006). An atlas of pollen-vegetation-climate relationships for the United States and Canada. *American Association of Stratigraphic Palynologists Foundation*.

Woodward, F. I. (1987). *Climate and plant distribution*. Cambridge University Press, Cambridge.

Wu, H., Guiot, J., Brewer, S., and Guo, Z. (2007a). Climatic changes in Eurasia and Africa at the last glacial maximum and mid-Holocene: reconstruction from pollen data using inverse vegetation modelling. *Climate Dynamics*, 29(2-3):211–229.

Wu, H., Guiot, J., Brewer, S., Guo, Z., and Peng, C. (2007b). Dominant factors controlling glacial and interglacial variations in the treeline elevation in tropical africa. *PNAS*, 104:9720–9724.

Wu, H., Guiot, J., Peng, C., and Guo, Z. (2009). New coupled model used inversely for reconstructing past terrestrial carbon storage from pollen data: validation of model using modern data. *Global Change Biology*, 15:82–96.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Zhang, H. and Wang, Y. (2009). Kriging and cross-validation for massive spatial data. *Environmetrics*, on-line.

Appendix A

Supplementary material chapter 2

A.1 Additional vegetation parameters

Tables of additional model parameters.

Species	$k_{la:sa}$ (m^2 m^{-2})	Leaf long. (y)	R_{fire}	Chilling (b, k)	Longevity (y)
Abies alba	6000	6	0.2	(100,0.05)	350
Alnus incana	6000	0.5	0.2	(100,0.05)	200
Betula pendula	6000	0.5	0.2	(500,0.02)	300
Betula pubescens	6000	0.5	0.1	(100,0.05)	300
Carpinus betula	6000	0.5	0.1	(1000,0.025)	150
Corylus avellana	5000	0.5	0.2	(200,0.05)	100
Fagus sylvatica	6000	0.5	0.1	(220,0.03)	400
Fraxinus excelsior	6000	0.5	0.1	(100,0.05)	400
Picea abies	6000	6	0.1	(100,0.05)	400
Pinus sylvestris	3500	2	0.4	(100,0.05)	500
Pinus halepensis	4000	2	0.4	(100,0.05)	350
Populus tremula	6000	0.5	0.2	(100,0.05)	160
Quercus coccifera	3200	3	0.5	(100,0.05)	350
Quercus ilex	4000	3	0.3	(100,0.05)	350
Quercus robur	6000	0.5	0.2	(100,0.05)	500
Tilia cordata	6000	0.5	0.1	(1000,0.025)	500
Ulmus glabra	6000	0.5	0.1	(100,0.05)	400
C3 grass	-	1	1.0	-	1

Table A.1: Additional species parameters and bioclimatic limits used in the LPJ-GUESS model. $k_{la:sa}$: ratio leaf area to sapwood cross-sectional area. Leaf long.: leaf longevity. R_{fire}: Fraction of a species' patch population and litter that survives a fire. Chilling (b, k): chilling parameters, as described by Sykes et al. (1996). Longevity: tree species longevity.

Shade Class	parff_{min} (10^5 J m^{-2} d^{-1})	greff_{min} (kg C_{leaf} m^{-2} y^{-1})	est_{max} (m^{-2} y^{-1})	α	conv_{sap} (y^{-1})
St	3.50	0.05	0.05	3	0.10
Ist	5.75	0.06	0.10	6	0.15
Si	8.00	0.07	0.30	9	0.20

Table A.2: Shade tolerance parameters used in the LPJ-GUESS model. See Smith et al. (2001) for full details. parff_{min}: Minimum photosynthetically active radiation at the forest floor for establishment. greff_{min}: Growth efficiency threshold. est_{max}: Maximum sapling establishment rate. α:recruitment shape parameter. conv_{sap}: Sapwood to hardwood conversion rate. Relative to Si species, St tree species require less photosynthetically active radiation at the forest floor to establish, produce fewer saplings under full light conditions, have a lower threshold growth efficiency for stress mortality, have less suppression of establishment at low forest-floor NPP, and convert proportionally less sapwood to hardwood annually. ISt species have intermediate characteristics.

A.2 Comprehensive description of the particle filter

The aim of this appendix is to fully describe the particle filter algorithm used for inference. We start from the model's equations and show how to obtain the posterior distribution.

A.2.1 Initialisation of the algorithm

For time t_1 of the first pollen core sample, according to the Bayes theorem and a development from Equation 2.5

$$p(V_{t_1}, C_{t_1}|Y_{t_1}) \propto p(Y_{t_1}|V_{t_1}).p_{\text{LPJ}}(V_{t_1}|V_{t_0}, C_{t_1}).p(C_{t_1})$$

Time t_0 (see Figure 2.3) is a notation to show that the time relationship between vegetation always exists. As we do not have data to reconstruct t_0 vegetation we assume an equilibrium hypothesis between climate and vegetation at time t_1 and then use:

$$p(V_{t_1}, C_{t_1}|Y_{t_1}) \propto p(Y_{t_1}|V_{t_1}).p_{\text{LPJ}}(V_{t_1}|C_{t_1}).p(C_{t_1})$$

From these equations the algorithm will be: Sample N_p "particles" $C_{t_1}^{(l=1:N_p)}$ from $p(C_{t_1})$. "Particles" describe the simulated climates and later the simulated couples of climate-vegetation. For each particle we run LPJ-GUESS for 500 years to reach equilibrium and obtain N_p "particles" $(C_{t_1}, V_{t_1})^{(l=1:N_p)}$ following $p_{\text{LPJ}}(V_{t_1}|C_{t_1}).p(C_{t_1})$. For each particle we then compute the non-normalized importance weights:

$$\omega_{t_1}^{(l)} = \frac{p(Y_{t_1}|V_{t_1}^{(l)}).p_{\text{LPJ}}(V_{t_1}^{(l)}|C_{t_1}^{(l)}).p(C_{t_1}^{(l)})}{p_{\text{LPJ}}(V_{t_1}^{(l)}|C_{t_1}^{(l)}).p(C_{t_1}^{(l)})} = p(Y_{t_1}|V_{t_1}^{(l)})$$

normalizing the weights (so their sum is 1) we obtain

$$\tilde{\omega}_{t_1}^{(l)} = \frac{\omega_{t_1}^{(l)}}{\sum_{k=1}^{N_p} \omega_{t_1}^{(k)}}$$

A discrete approximation of $p(V_{t_1}, C_{t_1}|Y_{t_1})$ is therefore

$$\tilde{p}(V_{t_1}, C_{t_1}|Y_{t_1}) = \sum_{l=1}^{N_p} \tilde{\omega}_{t_1}^{(l)}.\delta_{(V_{t_1}, C_{t_1})^{(l)}}$$

where $\delta_{(V_{t_1}, C_{t_1})}$ is the Dirac mass applied at (V_{t_1}, C_{t_1}).

A.2.2 Step t_j of the algorithm

Let t_i and t_j be two consecutive core times. Starting at t_j step of the algorithm we have $(V_{t_1:t_i}, C_{t_1:t_i})^{(l=1:N_p)}$, N_p "histories" of vegetation and climate weighted by $\tilde{\omega}_{t_i}^{(l=1:N_p)}$. These series and weights define the discrete approximation of $p(V_{t_1:t_i}, C_{t_1:t_i}|Y_{t_1:t_i})$. We want to add a coherent "particle" $(V_{t_j}, C_{t_j})^{(l)}$ to each history. Each new history obtained by concatenation of $(V_{t_1:t_i}, C_{t_1:t_i})^{(l)}$ and $(V_{t_j}, C_{t_j})^{(l)}$ with their associated weights $\tilde{\omega}_{t_j}^{(l)}$ must define the discrete approximation of $p(V_{t_1:t_j}, C_{t_1:t_j}|Y_{t_1:t_j})$.

By Bayes theorem and the model definition $p(V_{t_1:t_j}, C_{t_1:t_j}|Y_{t_1:t_j})$ can be developed

$$\begin{aligned} p(V_{t_1:t_j}, C_{t_1:t_j}|Y_{t_1:t_j}) &\propto p(Y_{t_j}|V_{t_j}).p(V_{t_j}, C_{t_j}|V_{t_i}).p(V_{t_1:t_i}, C_{t_1:t_i}|Y_{t_1:t_i}) \\ &\propto p(Y_{t_j}|V_{t_j}).p_{\mathrm{LPJ}}(V_{t_j}|V_{t_i}, C_{t_j}).p(C_{t_j}).p(V_{t_1:t_i}, C_{t_1:t_i}|Y_{t_1:t_i}) \end{aligned}$$

We simply have to sample N_p "particles" $C_{t_j}^{(l=1:N_p)}$ of climate parameters from $p(C_{t_j})$. For each particle we run LPJ-GUESS for $t_j - t_i$ years and obtain N_p "particles" $(C_{t_j}, V_{t_j})^{(l=1:N_p)}$ following $p_{\mathrm{LPJ}}(V_{t_j}|V_{t_i}, C_{t_j}).p(C_{t_j})$. For each particle we then recompute non-normalized importance weights:

$$\begin{aligned} \omega_{t_j}^{(l)} &= \frac{p(Y_{t_j}|V_{t_j}^{(l)}).p_{\mathrm{LPJ}}(V_{t_j}^{(l)}|V_{t_i}^{(l)}, C_{t_j}^{(l)}).p(C_{t_j}^{(l)}).p(V_{t_1:t_i}^{(l)}, C_{t_1:t_i}^{(l)}|Y_{t_1:t_i})}{p_{\mathrm{LPJ}}(V_{t_j}^{(l)}|V_{t_i}^{(l)}, C_{t_j}^{(l)}).p(C_{t_j}^{(l)})} \\ &= p(Y_{t_j}|V_{t_j}^{(l)}).\sum_{k=1}^{N_p} \tilde{\omega}_{t_i}^{(k)}.\delta_{(V_{t_1:t_i}, C_{t_1:t_i})^{(k)}} \\ &= p(Y_{t_j}|V_{t_j}^{(l)}).\tilde{\omega}_{t_i}^{(l)} \end{aligned}$$

and then renormalize the weights according to

$$\tilde{\omega}_{t_j}^{(l)} = \frac{\omega_{t_j}^{(l)}}{\sum_{k=1}^{N_p} \omega_{t_j}^{(k)}}$$

A.2.3 Regeneration

The sequential importance sampling algorithm presented above is theoretically valid, but its efficiency decreases with time, i.e. after a number of time steps the discrete approximation of the posterior distribution will be reduced to one single particle with

weight equal to 1. The solution is to add a regeneration step making the algorithm a particle filter algorithm (Doucet et al., 2001).

The regeneration step consists of sampling with replacement particles according to their weights. It implies that a particle with a weight of 0 will not be sampled further and is removed but that those with high weights will be sampled many times and are therefore multiplied. Since the regeneration step introduces a Monte Carlo error in the estimation, we do not have to do it if the particles are well distributed (Doucet et al., 2001). The criterion used to determine the need to resample the Effective Sample Size (ESS) criterion.

$$\mathrm{ESS}_t = \left(\sum\nolimits_{l=1}^{N_p} \left(\tilde{\omega}_t^{(l)} \right)^2 \right)^{-1}$$

The ESS criterion takes its values in the range 0 to N_p. If the degeneracy of particles is too high (and thus the ESS lies under an arbitrary threshold of $N_p/2$) we apply the regeneration step (resample) and we reset all weights to $1/N_p$, otherwise we keep all particles and weights.

Different methods are available for the resampling step. We use the efficient residual sampling technique from Liu and Chen (1998): At step t we have N_p particles $(V_t, C_t)^{(l=1:N_p)}$ weighted by $\tilde{\omega}_t^{(l=1:N_p)}$. In a first step, for each particle (l), we keep $n^{1,(l)} = \lfloor N_p \tilde{\omega}_t^{(l)} \rfloor$ copies of the particle. In a second step we randomly sample $m = N - \sum_{l=1}^{N_p} n^{1,(l)}$ particles in the set of all particles weighted by $\omega_t^{1,(l)} \propto N_p . \tilde{\omega}_t^{(l)} - n^{1,(l)}$.

Appendix B

Supplementary material chapter 3

B.1 Mean and variance of a Poisson ratio

Let $N \sim \mathcal{P}(\lambda)$ be a discrete random value with Poisson distribution and $\lambda \geq 0$. Let $M \sim \mathcal{P}(\mu)$ be another discrete random value, independent of N, with Poisson distribution and $\mu \geq 0$. Our interest lies in $R = \frac{N}{N+M}$, when $K = N + M > 0$ and in particular in $\mathbb{E}[R|K > 0]$ and $\mathbb{V}\text{ar}[R|K > 0]$.

We note that, conditional on $K = k$, $N \sim B(k, p)$ a binomial distribution with k outcomes and $p = \frac{\lambda}{\lambda+\mu}$.

Thus $\mathbb{E}[R|K = k] = p$ and $\mathbb{V}\text{ar}[R|K = k] = \frac{1}{k}p(1-p)$. From the former $\mathbb{E}[R|K > 0] = \mathbb{E}_{K>0}[\mathbb{E}[R|K]] = p$.

Since $\mathbb{E}[R^2|K] = p^2 + \frac{1}{K}p(1-p)$ we have $\mathbb{E}[R^2|K > 0] = p^2 + p(1-p)\mathbb{E}_{K>0}\left[\frac{1}{K}\right]$, where $K \sim P(\nu)$ and $\nu = \lambda + \mu$. But

$$\begin{aligned} \mathbb{E}_{K>0}\left[\frac{1}{K}\right] &= \frac{e^{-\nu}}{1-e^{-\nu}}\left(\sum_{k>0}\frac{\nu^k}{k!k}\right) \\ &= \frac{e^{-\nu}}{1-e^{-\nu}}\int_0^\nu \frac{e^u - 1}{u}\,du \end{aligned}$$

which is not analytically tractable.

We can obtain bounds for $\sum_{k>0} \frac{\nu^k}{k!k}$,

$$\begin{aligned}
\frac{1}{\nu}\left(\sum_{k=1}^{\infty}\frac{\nu^{k+1}}{(k+1)!}\right) &< \sum_{k>0}\frac{\nu^k}{k!k} &= \frac{1}{\nu}\left(\sum_{k=1}^{\infty}\frac{\nu^{k+1}}{(k+1)!}.\frac{k+1}{k}\right) \\
\frac{1}{\nu}\left(e^{\nu}-1-\nu\right) &< \sum_{k>0}\frac{\nu^k}{k!k} &= \frac{1}{\nu}\left(e^{\nu}-1-\nu+\sum_{k=1}^{\infty}\frac{\nu^{k+1}}{(k+1)!k}\right) \\
\frac{1}{\nu}\left(e^{\nu}-1-\nu\right) &< \sum_{k>0}\frac{\nu^k}{k!k} &< \frac{1}{\nu}\left(e^{\nu}-1-\nu+\frac{3}{\nu}\left(\sum_{k=1}^{\infty}\frac{\nu^{k+2}}{(k+2)!}\right)\right) \\
\frac{1}{\nu}\left(e^{\nu}-1-\nu\right) &< \sum_{k>0}\frac{\nu^k}{k!k} &< \frac{1}{\nu}\left(e^{\nu}-1-\nu\right)+\frac{3}{\nu^2}\left(e^{\nu}-1-\nu-\frac{\nu^2}{2}\right)
\end{aligned} \tag{B.1}$$

They bound $\mathbb{E}_{K>0}\left[\frac{1}{K}\right]$ to

$$\frac{1-e^{-\nu}-\nu e^{-\nu}}{\nu(1-e^{-\nu})} < E_{K>0}\left[\frac{1}{K}\right] < \frac{1-e^{-\nu}-\nu e^{-\nu}}{\nu(1-e^{-\nu})}+\frac{3(1-e^{-\nu}-\nu e^{-\nu}-\nu^2e^{-\nu}/2)}{\nu^2(1-e^{-\nu})}$$

Hence $\mathbb{E}_{K>0}\left[\frac{1}{K}\right] \approx \frac{1}{\nu}$ and $\mathbb{V}\mathrm{ar}\left[R|K>0\right] \approx \frac{1}{\nu}p(1-p)$ for ν 'sufficiently large'. This demonstrates that the model works like a Multinomial-Dirichlet model with one ν overdispersion parameter for the probabilities.

Appendix C

Supplementary material chapter 4

C.1 Pollen diagrams of the four studied sites

The pollen data used for the four sites are part of the European Pollen Database (`www.europeanpollendatabase.net`). They were published in Göransson (1991); Digerfeldt (1972, 1977); Almquist-Jacobson (1994). For linkage with the vegetation model outputs, we grouped the original taxa into 15 groups including the 14 major European trees and a group called 'GrSh' comprising the grasses and shrubs. For more information see Garreta et al. (2009). On the pollen diagrams we removed the Abies and the Quercus evergreen groups that are completely absent from the four studied cores.

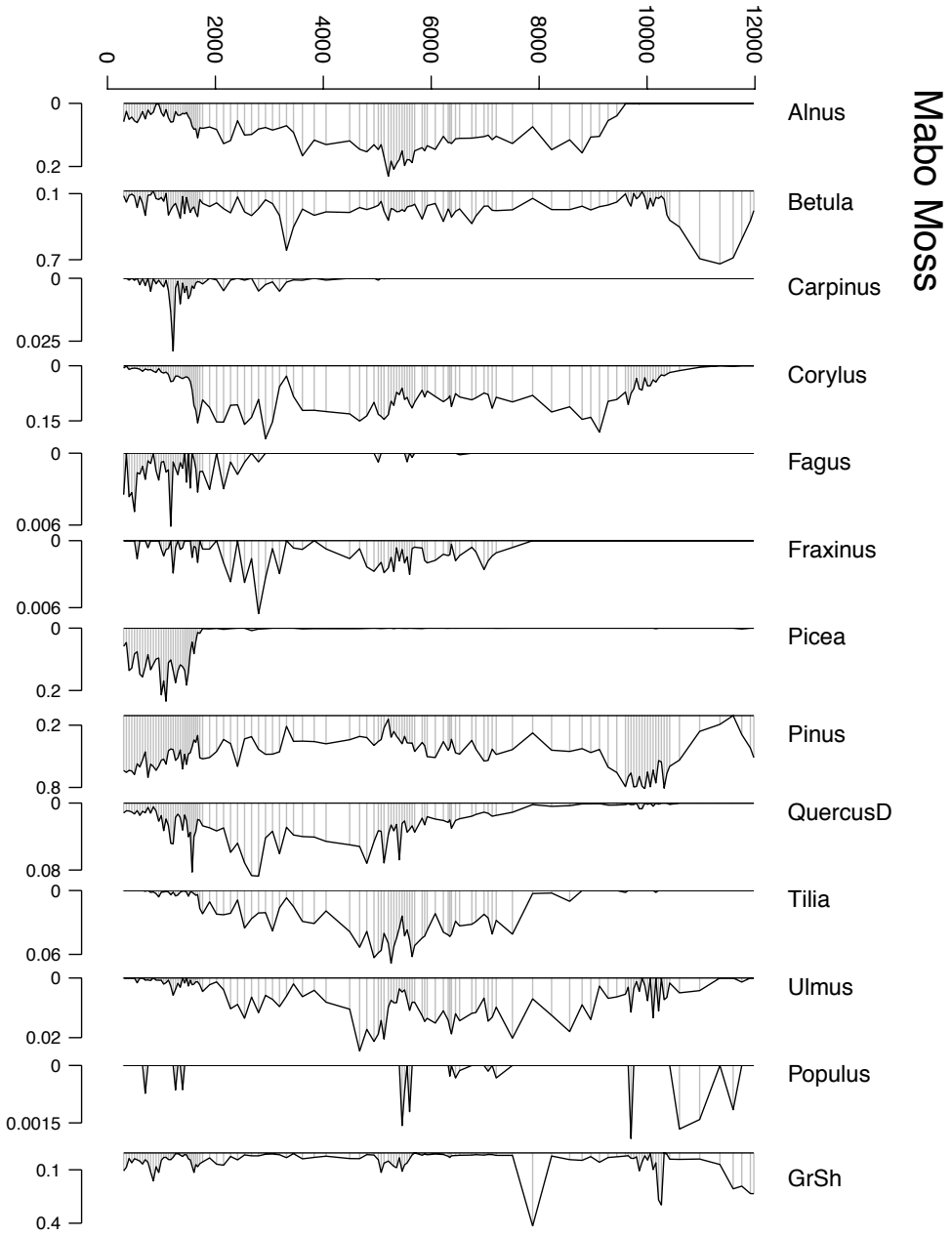

Figure C.1: Pollen diagram of the Mabo Moss sediment core (Göransson, 1991). (x-axis) Age in years before 1950. (y-axis) The pollen proportion per pollen taxa. **Caution** The scales are adjusted and hide very large differences in the composition.

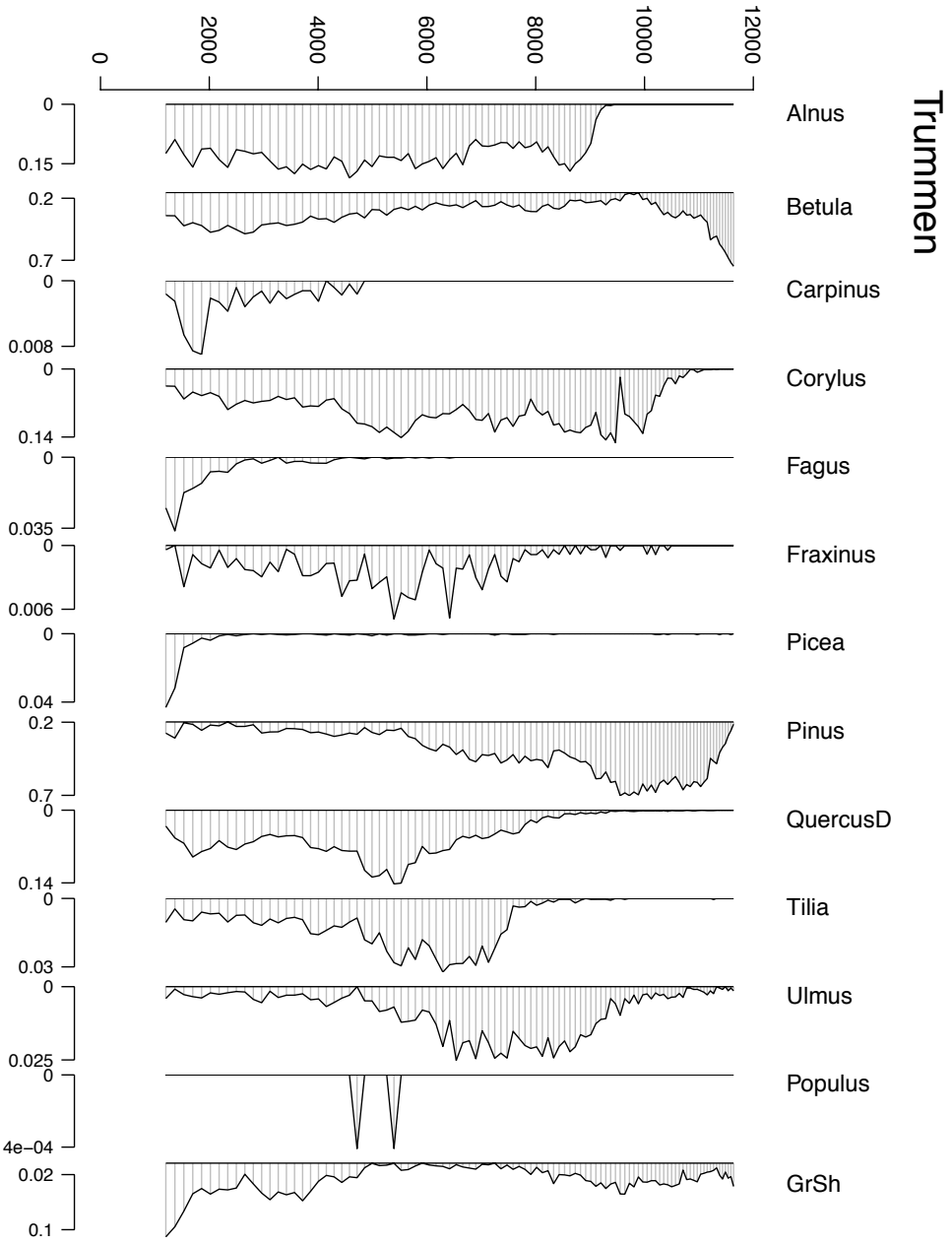

Figure C.2: Pollen diagram of the Trummen sediment core (Digerfeldt, 1972). (x-axis) Age in years before 1950. (y-axis) The pollen proportion per pollen taxa. **Caution** The scales are adjusted and hide very large differences in the composition.

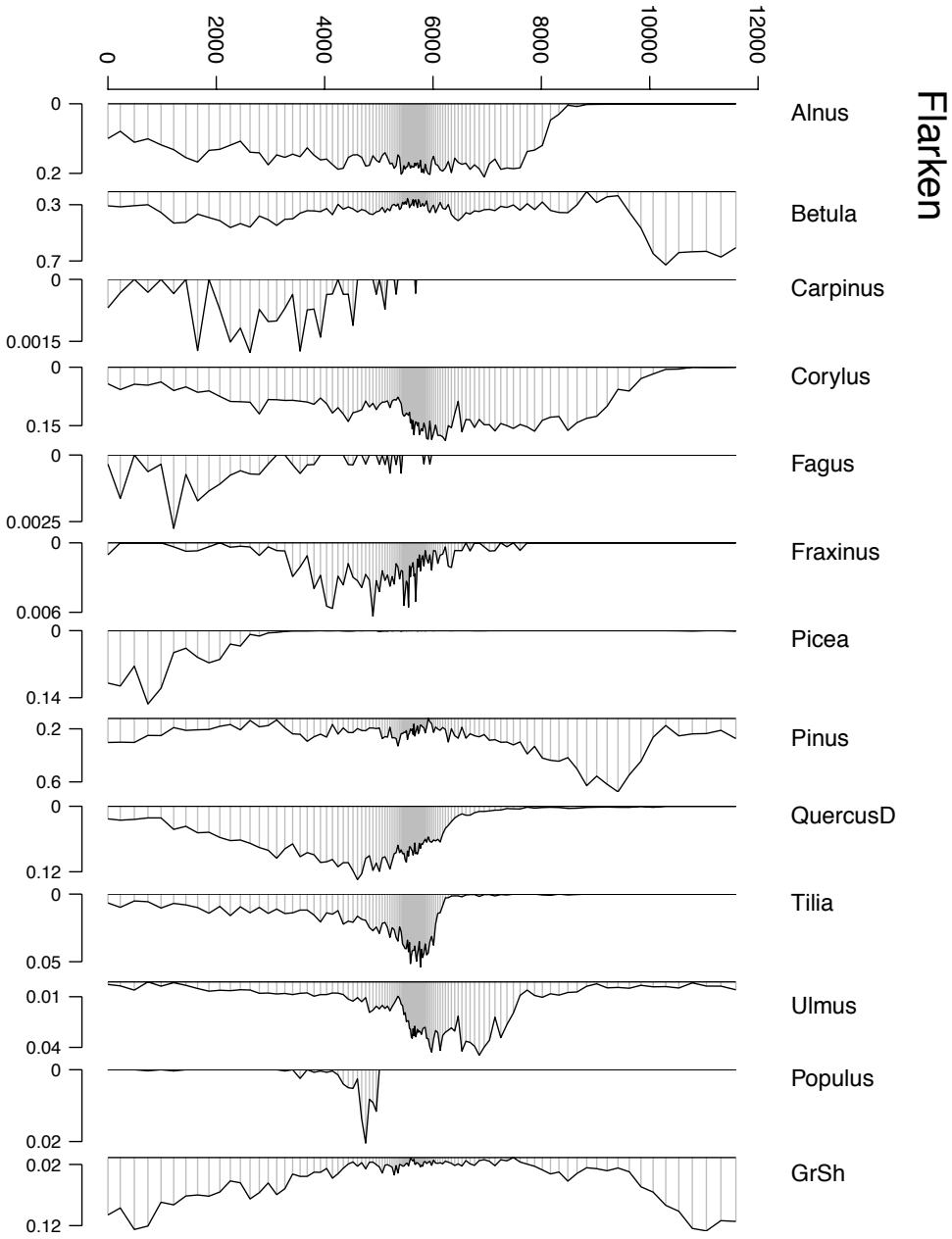

Figure C.3: Pollen diagram of the Flarken sediment core (Digerfeldt, 1977). (x-axis) Age in years before 1950. (y-axis) The pollen proportion per pollen taxa. **Caution** The scales are adjusted and hide very large differences in the composition.

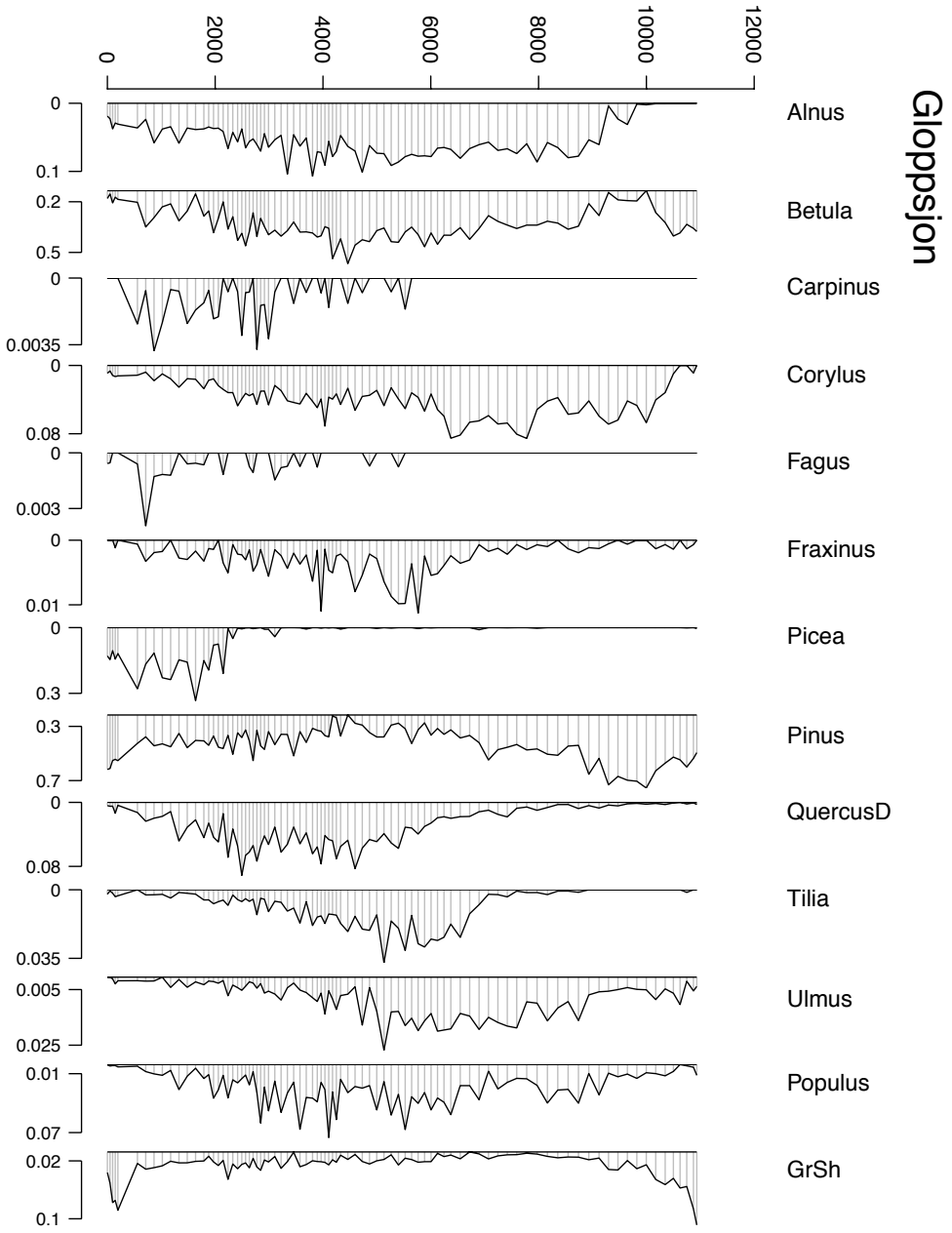

Figure C.4: Pollen diagram of the Gloppsjon sediment core (Almquist-Jacobson, 1994). (x-axis) Age in years before 1950. (y-axis) The pollen proportion per pollen taxa. **Caution** The scales are adjusted and hide very large differences in the composition.