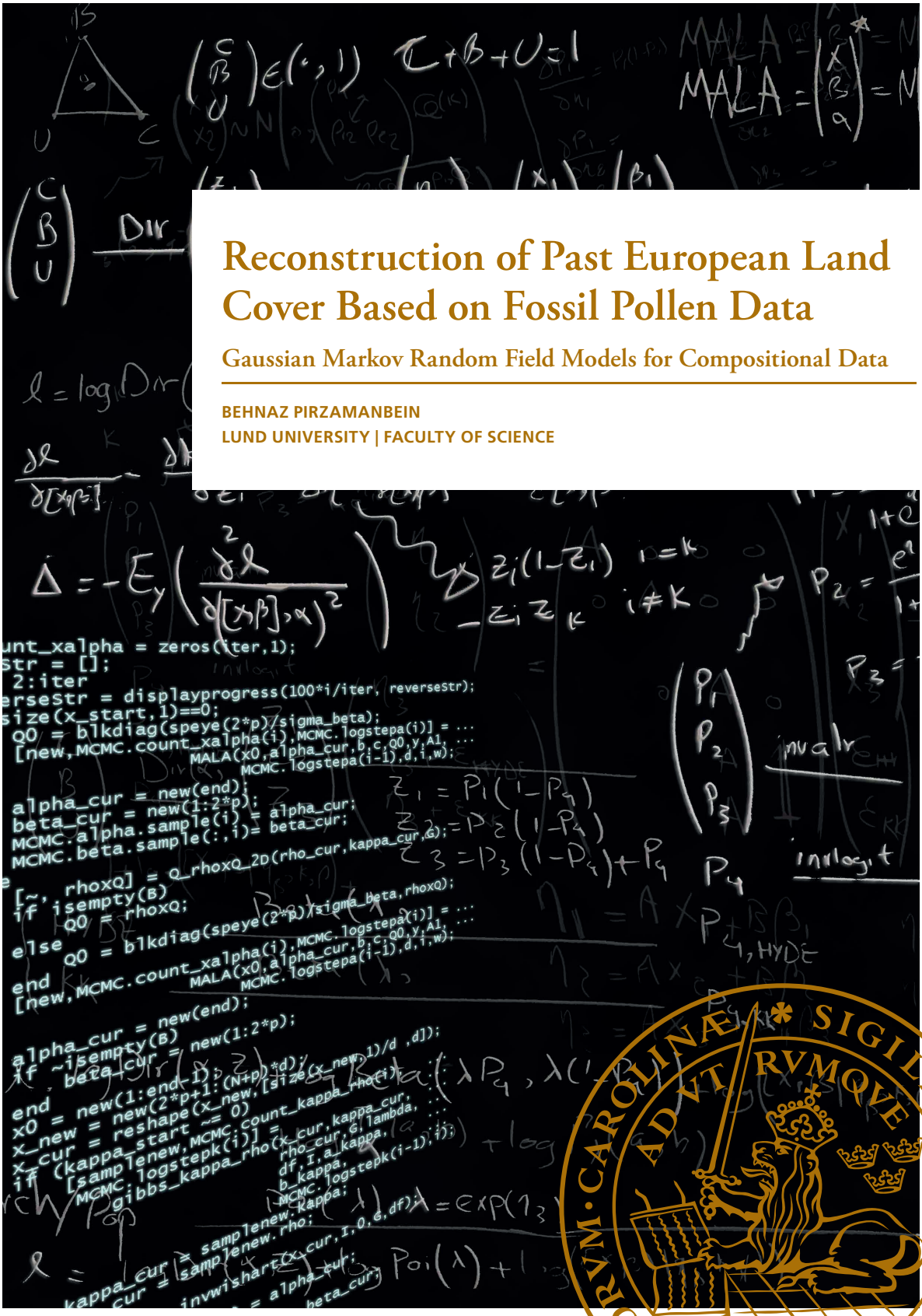
Reconstruction of Past European Land Cover Based on Fossil Pollen Data

Gaussian Markov Random Field Models for Compositional Data

Pirzamanbein, Behnaz



Reconstruction of Past European Land Cover Based on Fossil Pollen Data

Gaussian Markov Random Field Models for Compositional Data

BEHNAZ PIRZAMANBEIN
LUND UNIVERSITY | FACULTY OF SCIENCE

Reconstruction of Past European Land Cover Based on Fossil Pollen Data

Gaussian Markov Random Field Models for Compositional Data

Behnaz Pirzamanbein

Lund University

Faculty of Science
Centre for Mathematical Sciences
Centre for Environmental and Climate Research

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Sweden

http://www.maths.lu.se/

Doctoral Thesis 2016:12
ISSN 1404-0034

ISBN 978-91-7753-076-3 (print)
978-91-7753-077-0 (pdf)
LUSFMS-1022-2016

Cover by Y. B. Bergström, M. Millnert and B. Pirzamanbein

Printed in Sweden by Media Tryck, Lund 2016

Abstract

The aim of this thesis is to develop statistical models to reconstruct past land cover composition and human land use based on fossil pollen records over Europe for different time periods over the past 6000 years. Accurate maps of past land cover and human land use are needed when studying the interaction between climate and land surface, and the effects of human land use on past climate. Existing land cover maps are mainly simulations from dynamic vegetation models and anthropogenic land cover change scenarios. Pollen records is an alternative to existing land cover estimates that might give better insight into past land cover. The pollen counts are extracted from lake and bog sediments and used to estimate the three land cover compositions; coniferous forest, broadleaved forest, and unforested land for grid cells surrounding the lakes and bogs.

In this thesis, first, a statistical model is developed to interpolate transformed pollen based land cover compositions (PbLCC) with spatial dependency modelled using a Gaussian Markov random Field (GMRF). The mean structure is modelled using a regression on different sets of covariates including elevation and model based vegetation estimates. The model is fitted using Integrated Nested Laplace Approximation. The results indicated the existence of spatial dependence structure in the PbLCC and the possibility of reconstructing past land cover from PbLCC. If the compositional data is over-dispersed, the transformed Gaussian model might underestimate the uncertainties. To capture the variation in the composition correctly, a Bayesian hierarchical model (BHM) for Dirichlet observations of a GMRF is developed. The model is estimated using MCMC with sparse precision matrix of the GMRF being used for computational efficiency. Comparison between the Dirichlet and Gaussian models showed the advantages of the Dirichlet in describing the PbLCC. The large discrepancies in the model based estimates used as covariates could affect the Dirichlet models ability to reconstruct past land cover. To assess this concern a sensitivity study was performed, showing that the results are robust to the choice of covariates. Finally, the BHM is extended to reconstruct past human land use by combing the PbLCC with anthropogenic land cover change estimates. This extension aims at decomposing the PbLCC into past natural land cover and human land use.

Popular Summary

Spatial distribution of land cover plays an important role in climate system and global carbon cycle. Research shows that changes in land cover are associated with large climatic effects. These changes are either due to climate change or human activities. Human can influence and change the abundance of land cover through deforestation, urbanization and agriculture. Studies show that replacing forests with agricultural land decreases the temperature while urbanization causes local increases in temperature. Comparing the historical temperature records with past natural and human induced land cover might give a better understanding of the interactions among climate, land cover and human effects.

The problem is the existence of considerably different descriptions of past land cover and human land use. Existing land cover descriptions are based on natural land cover combined with human land use. Past human land use maps are mainly based on simulations of human population density and the amount of agricultural land needed to feed the given population. Furthermore, natural land cover maps are simulations based on past climate including temperature, precipitation and soil type; they represent the natural vegetation that can grow in certain climate conditions without considering human activity. The differences in these available maps are caused by differences in the model assumptions, as well as the simulations of climate variables and population density.

On the other hand, fossil pollen counts can be used to estimate past land cover based on local observations over the past 10 000 years. The only problem is that the information on pollen counts, extracted from lakes and bogs, are limited in reproducing the land cover for the area surrounding these lakes and bogs.

This thesis aims to develop statistical models that can create continuous maps of past land cover and human land use based on pollen observations.

Since the spread of pollen as well as certain climate conditions lead to the growth of similar types of vegetation within a spatial range, one can expect to observe similar vegetation types in areas closer to each other than farther apart. Because of this fact, spatial statistics is used as a main tool to identify and model this space dependency in the pollen observations.

Acknowledgements

I would like to thank everyone who consciously or unconsciously helped me during my PhD studies.

Special thanks to my supervisor Johan Lindström for leading me on this path with all his help, support and patience; without his guidance and encouragement this work would not have been possible.

I would like to thank my co-authors, Marie-José and Anneli for their support, useful discussions, enthusiasm, and for making me believe that what I do matters. I would like to thank all my colleagues in mathematical statistics, past and present, for all the interesting discussions during the breaks and for Fridays' fika. My special thanks to Umberto for introducing me to the student paper competition, which ended in great success, and to James, for always making us smile. I would like to thank Lise-Lotte, Mona and Maria for their unlimited support and for making me feel at home. And a special thanks to my colleagues at CEC and especially Pål-Axel and Markku.

I would like to thank my friends: Yalda for her unlimited kindness, Kimia for her infinite soul and for saving me from my logical world, Setayesh for being my idol and older sister, Fioralba for her limitless positive energy and love, Ala for the encouragements before and during my PhD, Anu and Nicola for their support both in my Badminton life and on my path of becoming a scientist, Sahar, Saira, Sara Sh for always being there though far away, and all LUGI Badminton friends for the joyful moments we create together, which gives me a fresh start everyday.

I would like to thank my third grade primary school teacher, Mrs Jahangiri for planting the seed of my greatest dream: becoming a scientist.

I would like to thank my parents for their priceless love and support, my father, Anoush, for being an outlier and teaching me to stand for what's right, my mother, Hamideh, for being the strongest woman I have met in my life and for being my hero, my family, especially my grandma, my aunt Maryam, Nahid, Hamid, and my brother Behrang.

I would like to thank Martin, for his infinite love and pure heart, his patience and wisdom, all the emotional and nerdy support, and for being there to bina-hayat and abadiyat.

Lund, November 2016 *Behnaz Pirzamanbein*

Contents

List of papers

This thesis consists of the following papers:

Paper A Behnaz Pirzamanbein, Johan Lindström, Anneli Poska, Shinya Sugita, Anna-Kari Trondman, Ralph Fyfe, Florence Mazier, Anne B. Nielsen, Jed O. Kaplan, Anne E. Bjune, H. John B. Birks, Thomas Giesecke, Mikhel Kangur, Małgorzata Latałowa, Laurent Marquer, Benjamin Smith and Marie-José Gaillard: Creating spatially continuous maps of past land cover from point estimates: A new statistical approach applied to pollen data
Ecological Complexity, 2014:12, 20, 127–141.

Paper B Behnaz Pirzamanbein, Johan Lindström, Anneli Poska, and Marie-José Gaillard: Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties.
Published as preprint on Arxiv: `arxiv.org/abs/1511.06417`
The paper was awarded for **Outstanding Bayesian research applied to climate science** in the student paper competition of the Section on Bayesian Statistical Science (SBSS), American Statistical Association (ASA), 2016.
The paper was also recognized with an **Honorary Mention** in the student paper competition of the Section on Statistics and the Environment (ENVR), ASA, 2016.
Both at the joint statistical meeting 2016.
To be submitted to *Journal of the American Statistical Association*.

Paper C Behnaz Pirzamanbein, Anneli Poska, Johan Lindström: Analysing the sensitivity of pollen based land cover maps to different auxiliary variables
To be submitted to *Climate of Past*.

Paper D Behnaz Pirzamanbein, Johan Lindström: Reconstruction of past Human Land Use from Pollen data and Anthropogenic Land Cover Changes scenarios
To be submitted to *Environmetrics*.

Introduction

In climate studies the spatial distribution of past vegetation/land cover plays an important role (Claussen et al., 2001). To characterize the feedbacks of climate and earth surface (vegetation/land cover) and assess the influence of human land use on past climate, precise estimates of past land cover are required (Gaillard et al., 2010, Strandberg et al., 2014).

In climate models and studies of past climate, three fractions of land cover are often used; coniferous forest, broadleaved forest and unforested land. The latter, unforested land, can further be decomposed into natural openness and human land use. Past land cover used in climate models are often based on potential natural vegetation obtained from dynamic vegetation models (DVMs) (e.g. LPJ-GUESS, Smith et al., 2001). The DVM simulations use climatic variables such as, temperature, precipitation, and soil type to produce natural vegetation cover without considering the influences of human activity on land cover. Therefore, the potential natural vegetation is combined with anthropogenic land cover changes scenarios (ALCC; e.g. Kaplan et al., 2009, Klein Goldewijk et al., 2011) to create more realistic land cover estimates. The ALCCs produce estimates of human land use mainly based on human population estimates.

As an alternative, fossil pollen records provide observation based descriptions of past vegetation, including effects of human land use, during the entire Holocene (approx. 10 000 BCE to present day) (Gaillard et al., 2010). The pollen counts are extracted from lake and bog sediments and only represent vegetation in the area surrounding each studied sites. Further, past land cover composition based on fossil pollen counts can be estimated using the REVEALS model (Regional Estimates of VEgetation Abundance from Large Sites) for limited number of sampling sites (Sugita, 2007). In order to use the pollen based land cover compositions in climate models or when studying past climate, interpolation from the sampling sites to continues maps of past land cover at regional and subcontinental scales are needed.

Statistical modelling of species compositions is a common problem in environmental studies. For example, Billheimer et al. (2001) modelled trophic compositions in order to investigate the stability of arthropod communities in the

presence of environmental disturbance, Tjelmeland and Lund (2003) modelled sediments compositions in an Arctic lake and Paciorek and McLachlan (2009) used pollen records to recover forest composition in the USA. Using spatial statistics, one can identify the spatial dependences between the data points. These dependencies can be used to reconstruct continuous maps of the land cover compositions.

The aim of this thesis is to reconstruct the past land cover and human land use proportions based on fossil pollen records in North Europe for different time periods over the past 6000 years. A Bayesian hierarchical model is developed with fast inference methods to create a flexible statistical model to produce new, observation based, land cover maps for climate modellers, palaeoecologists, and researchers studying the past climate.

The outline of this thesis is as follows: in Section 1, a brief descriptions of the pollen based land cover data, study domain and study time periods are given. Section 2 provides the definition and describes the important properties of compositional data. In Section 3, a summary of the statistical models is given. This section also includes developed models, inference and posterior uncertainties for composition. Finally, Section 4 concludes with summary of the papers included in this thesis.

1 Data

1.1 Pollen Data

The pollen based land cover data used for the reconstruction of past land-cover composition consists of three land cover types: Coniferous forest, Broadleaved forest and Unforested land. This data was obtained using the REVEALS model (Sugita, 2007). Trondman et al. (2015) applied REVEALS to 636 pollen records from lakes and bogs; producing estimates of regional land cover for 25 plant taxa which were then grouped into the 3 land cover types (see Table 1 and Appendix S2 in Trondman et al., 2015, for details). The regional estimates from REVEALS were obtained for $1° \times 1°$ grid cells (roughly 111.2×111.2 km^2), with each estimate accounting for all lakes and bogs within that grid cell (Hellman et al., 2008, Trondman et al., 2015).

REVEALS is a mechanistic model that takes into account the size of sedimentary basins and inter-taxonomic differences in pollen productivity and dispersal to estimate regional vegetation cover from fossil pollen records.

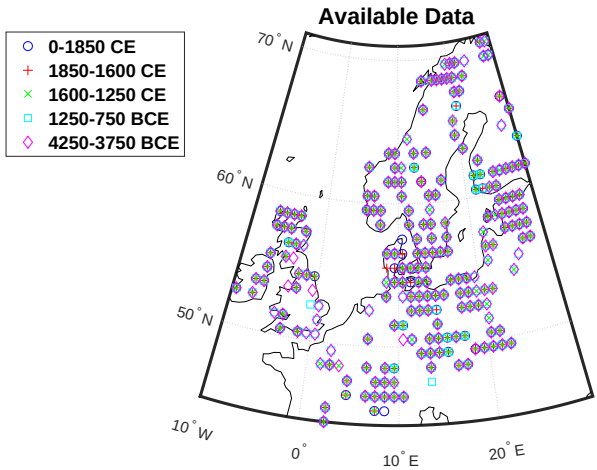


Figure 1: Available data for five time periods with 175, 181, 193, 204 and 196 observed grid cells

The study area covers Northwest and Western Europe and north of the Alps. The region has been divided into a spatial grid of $1° \times 1°$. The pollen based REVEALS estimates are available for five time periods that represent major climatic

and historical periods and are commonly used in both climate modelling and palaeoecological studies (Gaillard et al., 2010, Strandberg et al., 2014, Trondman et al., 2015)

Present–1850 CE: Recent past or industrial time with large human activity in Europe; this time period is also used to validate the modelling ("Present" means the most recent pollen records recovered at each site),

1850–1600 CE: End of the little Ice Age with a cold period in Europe, it is stated as pre-industrial time,

1600–1250 CE: Middle ages, this is the end of the Medieval warm period with an unstable period in European history with decreased human land use due to long periods of war and diseases, including the Black Death,

750–1250 BCE: Late Bronze Age transition with relatively high human impact in Europe, and

3750–4250 BCE: Early Neolithic with little human impact.

1.2 Human land use data

The human land use data is based on anthropogenic land cover changes scenarios. These scenarios are based on estimations of past human population densities, land area needed for food production to sustain that population, combined with a model of soil and climate suitability for food production. In this thesis two different anthropogenic land cover changes scenarios are used; 1) the KK10 scenarios of Kaplan et al. (2009), and 2) the History Database of the Global Environment (HYDE; Klein Goldewijk et al., 2011). The KK10 and HYDE data differ substantially for some of the past time periods due to differences in assumptions, modelling approaches, and historical records used. The major difference is that the KK10 scenario includes technology level as a component when estimating soil productivity while HYDE puts more weights on historical statistics of croplands and pastures.

2 Compositional data

In this section, a summarized description of compositional data and important properties that are used in this thesis are given.

If $\mathbf{y} = (y_1, y_2, \cdots, y_k)$ is a vector representing one observation of D-compositional data, then the compositional property implies

$$y_k \in (0,1), \qquad \text{and} \qquad \sum_k y_k = 1.$$

In order to model compositional data a transformation is often used to change the support from $(0,1)^D$ to $\mathbb{R}^D$ or $\mathbb{R}^{D-1}$. This is due to the fact that naive modelling of the untransformed data will not guarantee that the results remain in $(0,1)^D$. To achieve results in $(0,1)^D$, difficult conditions and constrains are needed for the untransformed data. Transforming to $\mathbb{R}$ allows for unconstrained modelling followed by a suitable inverse transformation of the results. One of the commonly used transformations (e.g. Aitchison, 1986, Billheimer et al., 2001, Tjelmeland and Lund, 2003) is the additive log ratio (alr)

$$\begin{aligned} \mathbf{u} &= g(\mathbf{y}) \qquad g : (0,1)^D \to \mathbb{R}^d \\ u_k &= \log \frac{y_k}{y_D}, \quad \text{for } k = 1, \cdots, d \end{aligned} \tag{1}$$

where $d = D - 1$, with the inverse transformation ($f = g^{-1}$)

$$\begin{aligned} \mathbf{u} &= f(\mathbf{y}) \qquad f : \mathbb{R}^d \to (0,1)^D \\ y_k &= \begin{cases} \dfrac{\exp(u_k)}{1 + \sum_k \exp(u_k)}, & \text{for } k = 1, \cdots d \\ \\ \dfrac{1}{1 + \sum_k \exp(u_k)}, & \text{for } k = D. \end{cases} \end{aligned} \tag{2}$$

The advantage of using alr compared to other transformation is that the alr decreases the dimensionality from D to $d = D - 1$ avoiding un-identifiable models. As a contrast, the central log-ratio transformation (clr; Aitchison, 1986) maps $(0,1)^D$ to $\mathbb{R}^D$, giving a $\mathbf{u}$ that is invariant to the addition of a constant (see e.g. Paciorek and McLachlan, 2009).

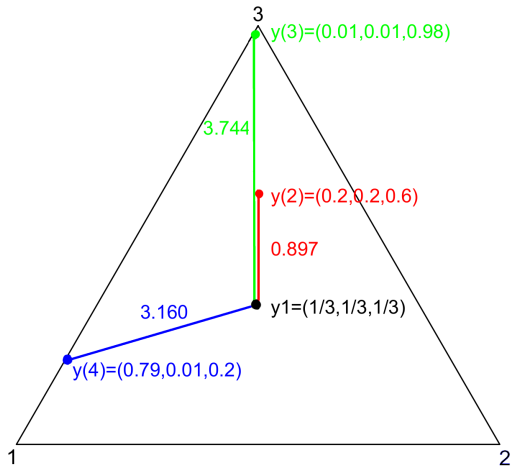


Figure 2: Ternary diagram of four different 3-compositional data points; y(1) = $(1/3, 1/3, 1/3)$ or the origin, y(2) = $(0.2, 0.2, 0.6)$, y(3) = $(0.01, 0.01, 0.98)$ and y(4) = $(0.79, 0.01, 0.2)$ with their distances to the origin (Figure 1 in Paper A).

The distance between two D-composition $\mathbf{U}$ and $\mathbf{V}$ is computed using the compositional distances,

$$\begin{aligned} \Delta(\mathbf{U},\mathbf{V}) &= \left[\sum_{i=1}^{D}\left(\log\frac{U_i}{\xi(\mathbf{U})} - \log\frac{V_i}{\xi(\mathbf{V})}\right)\right]^{1/2} \\ &= \left[(\mathbf{u}-\mathbf{v})^\top \mathbf{H}^{-1}(\mathbf{u}-\mathbf{v})\right]^{1/2} = \Delta(\mathbf{u},\mathbf{v}) \end{aligned} \tag{3}$$

where ξ is the geometric mean, $\xi(\mathbf{U}) = \sqrt[D]{U_1 U_2 \cdots U_D}$, $\mathbf{u}$ and $\mathbf{v}$ are alr transformations of the compositions $\mathbf{U}$ and $\mathbf{V}$ and $\mathbf{H}_{d\times d}$ is a matrix with elements $h_{ii} = 2$ and $h_{ij} = 1$ which neutralizes the choice of denominator in alr transformation. Similar to root mean square error, the average compositional distances is used for model evaluation. Figure 2 shows a simple example of compositional data and their distances. Note that getting closer to the edges and corners of the triangle, i.e. getting close to zero or one in one of the components, has higher impact on the distances than being around the origin $(1/3, 1/3, 1/3)$.

3 Spatial Statistics

Spatial statistics is a modern statistical field which analyses spatially distributed data, their dependencies and uncertainties. Spatial data are typically correlated in space so that the observations which are close to each other are more similar than the observations further apart.

3.1 Gaussian Random Fields

When modelling spatial data the main interest is often to use data, $\mathbf{y}$, at observed locations $\{\mathbf{s}_i\}_{i=1}^N$ to reconstruct/predict missing values at unobserved locations $\mathbf{s}_0$. A basic model for spatial data is to consider observations, $\mathbf{y}$, as being from an underlying stochastic (random) field, $\mathrm{x}(\mathbf{s})$, $\mathbf{s} \in \Omega \subset \mathbb{R}^2$. If all possible subsets of points in the field, $\mathrm{x}(s_1), \cdots, \mathrm{x}(s_n)$ are jointly multivariate Gaussian, the field is called a Gaussian Random Field with density,

$$\mathsf{P}(\mathbf{x}|\mu_{\mathrm{x}}, \mathbf{\Sigma}_{\mathrm{x}}) = \frac{1}{|2\pi\mathbf{\Sigma}_{\mathrm{x}}|^{\frac{1}{2}}} \exp\left((\mathbf{x}-\mu_{\mathrm{x}})^\top \mathbf{\Sigma}_{\mathrm{x}}^{-1}(\mathbf{x}-\mu_{\mathrm{x}})\right).$$

The statistical properties of a Gaussian random field can be completely specified by its mean function, $\mu_{\mathrm{x}}(\mathbf{s}) = \mathsf{E}(\mathrm{x}(\mathbf{s}))$, and covariance function, $\mathsf{C}(\mathbf{s},\mathbf{t}) = \mathsf{C}(\mathrm{x}(\mathbf{s}), \mathrm{x}(\mathbf{t}))$ with elements in $\mathbf{\Sigma}_{\mathbf{x}}$ given by $\mathsf{C}\left(s_i, t_j\right)$. The mean function captures the general trend in the random field and is commonly modelled using a regression on known functions (covariates) within the spatial coordinates. The spatial covariance function captures the dependency among the points in the field. If the mean is constant and covariance only depends on a vector between points; $\mathsf{C}(\mathbf{s},\mathbf{t}) = \mathsf{C}(\mathbf{s}-\mathbf{t})$, the field is said to be (weakly) stationary. A stationary field is further called isotropic if the covariance depends only on the distance between the points and not on the direction; $\mathsf{C}(\mathbf{s},\mathbf{t}) = \mathsf{C}\left(\|\mathbf{s}-\mathbf{t}\|\right)$. To obtain a stationary and isotropic covariance, a functional form of isotropic covariance is often used (see e.g. page 23 in Gelfand et al., 2010, for a list of different covariance functions). Among isotropic covariance function the most popular one, and commonly used in spatial statistics due to its flexible parametric form, is the Matérn covariance function:

$$\mathsf{C}(\mathbf{s},\mathbf{t}) = \sigma^2 \frac{(\kappa\|\mathbf{h}\|)^\nu K_\nu(\kappa\|\mathbf{h}\|)}{\Gamma(\nu)2^{\nu-1}} \tag{4}$$

where $\mathbf{h} = \|\mathbf{s}-\mathbf{t}\|$ is the distance between two points, ν controls the smoothness of the field, κ controls the range of dependency (range $= \sqrt{8\nu}/\kappa$), $\sigma^2 = \mathsf{C}(0)$ is

the field variance, Γ is a gamma function and $K_\nu(\cdot)$ is a modified Bessel function of the second kind.

3.2 Bayesian Hierarchical Model

The spatial fields are often used as components in Bayesian hierarchical models (Wikle, 2011). A hierarchical model is based on a joint distribution of all quantities in the model, specified through a series of conditional distributions. The hierarchical model typically consists of three parts:

Data model, $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, or likelihood which specifies the distribution of the measured/collected data given the underlying process and parameters,

Process model, $\pi(\mathbf{x}|\boldsymbol{\theta})$ which describes how the latent field behaves, and

Parameter model, $\pi(\boldsymbol{\theta})$, defining any prior knowledge regarding the parameters.

The inference for hierarchical model is based on the posterior distribution. Using Bayes' formula, the joint posterior is obtained by

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

In addition, the posterior of the parameters is obtained by marginalizing over $\mathbf{x}$

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \int \pi(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})\mathrm{d}\mathbf{x}, \tag{5}$$

while marginalizing over the parameters gives the posterior distribution of $\mathbf{x}$ as

$$\pi(\mathbf{x}|\mathbf{y}) \propto \int \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}. \tag{6}$$

Usually the posterior mean, $E(\mathbf{x}|\mathbf{y})$ is reported as the reconstruction of the latent field given observations $\mathbf{y}$, and the posterior variance $V(\mathbf{x}|\mathbf{y})$ is used as a measure of reconstruction uncertainties.

A common approach to modelling compositional data is to first apply a log-ratio transformation to the observations, so that $\mathbf{u} = g(\mathbf{y})$ using (1), and then model $\{u(s_i)\}_{i=1}^N$ using an underlying Gaussian field, $\mathbf{x}$, with additive Gaussian noise, $\boldsymbol{\epsilon} \sim \mathsf{N}\left(0, \mathbb{I}\sigma_\epsilon^2\right)$. Giving the hierarchical model as

$$\begin{aligned} \mathbf{u}(s_i)|\mathbf{x}(s_i), \sigma_\epsilon &\sim \mathsf{N}\left(\mathbf{x}(s_i), \mathbb{I}\sigma_\epsilon^2\right) \qquad i = 1, \cdots N \\ \mathbf{x}|\boldsymbol{\theta} &\sim \mathsf{N}\left(\boldsymbol{\mu}_{\mathrm{x}}(\boldsymbol{\theta}), \boldsymbol{\Sigma}_{\mathrm{x}}(\boldsymbol{\theta})\right). \end{aligned} \tag{7}$$

However, if the compositional data is over-dispersed, this modelling might underrate the uncertainty associated with the compositions (Paciorek and McLachlan, 2009). Another possibility is to assume that the data are observations from a Dirichlet distribution given the transformed underlying field (e.g. $\mathbf{z} = f(\mathbf{x})$ using (2)) and concentration parameter α to account for the dispersion in the composition. Resulting in the model

$$\begin{aligned} \mathbf{y}(s_i)|\mathbf{x}(s_i), \alpha &\sim \mathsf{Dir}\left(\alpha, f(\mathbf{x}(s_i))\right) \qquad i = 1, \cdots N \\ \mathbf{x}|\boldsymbol{\theta} &\sim \mathsf{N}\left(\mu_{\mathrm{x}}(\boldsymbol{\theta}), \boldsymbol{\Sigma}_{\mathrm{x}}(\boldsymbol{\theta})\right) \end{aligned} \tag{8}$$

were the Dirichlet density is

$$\mathsf{P}(\mathbf{y}|\mathbf{z}, \alpha) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha\, \mathrm{z}_k)} \prod_k \mathrm{y}_k^{\alpha\, \mathrm{z}_k - 1}, \quad \alpha > 0 \text{ and } \mathrm{z}_k > 0.$$

3.3 Computational issues and Gaussian Markov Random Fields

To compute the posterior (6) for both models, a Gaussian density needs to be evaluated. This requires the calculation of the inverse covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{x}}^{-1}$, and the log-determinant, $\log|\boldsymbol{\Sigma}_{\mathrm{x}}|$. For large geographic domains both of these computations are very expensive. Different methods exists to reduce the computational burden, such as, covariance tapering (Furrer et al., 2006) where small values in the covariance matrix are set to zeros creating a sparse covariance matrix, or low rank approximation methods including, fixed rank kriging (Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008), and process convolution (Higdon, 1998). The low rank methods are based on reduced basis function ,ψ, representations of the process model $\mathbf{x}$,

$$\mathrm{x}(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s}) \mathrm{w}_i, \qquad r \ll N \tag{9}$$

where r is predefined and fixed and $\mathbf{w} \sim \mathsf{N}(0, \boldsymbol{\Sigma}_{\mathrm{w}})$ which leads to a reduced size, $r \times r$, covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{w}}$.

An alternative and a good choice for modelling a Gaussian random field is a Gaussian Markov Random Field (**GMRF**). If a GRF $\mathbf{x}$, follows the Markov property, i.e. the distribution of x_i given the rest of the field is equal to the distribution of x_i given just the neighbours, then $\mathbf{x}$ is said to be a GMRF. The GMRF is parametrized using the inverse covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = \mathbf{Q}$, also called the

precision matrix. Because of the Markov property $\mathbf{Q}$ becomes sparse with all non-neighbour elements being zero. Thus, by choosing a small set of neighbours, $\mathbf{Q}$ contains many zeros. The sparse $\mathbf{Q}$ reduces the computational cost regarding the calculation of the inverse and determinant matrices compare to a dense covariance matrix.

To use a GMRF one needs to construct a useful and sparse $\mathbf{Q}$ matrix. Lindgren et al. (2011) developed a general method to construct GMRFs that approximate continuous GRFs with Matérn covariance. The method is based on the fact that GRFs with Matérn covariance function on $\mathbb{R}^d$ are solutions to the Stochastic Partial Differential Equation (**SPDE**)

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}\mathbf{x}(\mathbf{s}) = \tau\mathcal{W}(\mathbf{s}) \tag{10}$$

where $\mathcal{W}(\mathbf{s})$ is Gaussian white noise and Δ is the Laplace operator (Matérn, 1960, Whittle, 1954, 1963). The α is linked to the smoothness in (4) as $\alpha = \nu + \frac{d}{2}$ where d is the dimension of the domain on which the field is defined, and τ relates to the field variance through,

$$\sigma^2 = \mathsf{C}(0) = \frac{\tau^2\Gamma(\nu)}{(4\pi)^{\frac{d}{2}}\Gamma(\nu + \frac{d}{2})\kappa^{2\nu}}.$$

The main idea is to approximating $\mathrm{x}(\mathbf{s})$ as the solution to the SPDE in (10) using a basis expansion, $\mathrm{x}(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\mathrm{w}_i$. More specifically, a finite element solution is obtained by finding the distribution of the weights, $\mathbf{w}$, that fulfils the stochastic weak formulation of the SPDE. The weak formulation requires the following equality in distribution to hold for a specific set of test function $\varphi_j(\mathbf{s})$,

$$\langle\varphi_j, (\kappa^2 - \Delta)^{\frac{\alpha}{2}}\mathbf{x}\rangle \overset{d}{=} \langle\varphi_j, \tau\mathcal{W}\rangle, \qquad j = 1, \cdots, r \tag{11}$$

where $\langle f, g\rangle = \int f(\mathbf{s})g(\mathbf{s})\mathrm{d}\mathbf{s}$. Here, a simplified solution sketch for the case $\alpha = 2$ with Galerkin test functions $\psi_i = \varphi_i$ is shown (see Appendix C. in Lindgren et al., 2011, for details). Using the basis expansion, the left hand side of (11) can be written as $\mathbf{Kw} = \sum_{i=1}^{r} \mathrm{w}_i\langle\psi_j, (\kappa^2 - \Delta)\psi_i\rangle$ where $\mathbf{K}$ is a finite difference approximation matrix with elements

$$\mathbf{K}_{ji} = \langle\psi_j, (\kappa^2 - \Delta)\psi_i\rangle = \kappa^2\langle\psi_j, \psi_i\rangle + \langle\nabla\psi_j, \nabla\psi_i\rangle.$$

Defining $\mathbf{C}_{ji} = \langle\psi_j, \psi_i\rangle$ and $\mathbf{G}_{ji} = \langle\nabla\psi_j, \nabla\psi_i\rangle$ result in $\mathbf{K} = \kappa^2\mathbf{C} + \mathbf{G}$. The right hand side of (11) is a Gaussian vector with mean zero and covariance matrix

$$\mathsf{C}(\,\langle\psi_j, \tau\mathcal{W}\rangle, \langle\psi_i, \tau\mathcal{W}\rangle\,) = \tau^2\langle\psi_j, \psi_i\rangle = \tau^2\mathbf{C}.$$

Hence, the SPDE (in matrix form) becomes

$$\mathbf{Kw} \overset{d}{=} \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathsf{N}\left(0, \tau^2\mathbf{C}\right) \tag{12}$$

and $\mathsf{V}(\mathbf{w}) = \tau^2\mathbf{K}^{-1}\mathbf{C}\mathbf{K}^{-T}$. Therefore, $\mathbf{w} \sim \mathsf{N}\left(0, \mathbf{Q}^{-1}\right)$ with $\mathbf{Q} = \frac{1}{\tau^2}\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-\mathbf{1}}\mathbf{K}$ is a solution to (10).

The $\mathbf{Q}$ is sparse if $\mathbf{K}$ and $\mathbf{C}^{-1}$ are sparse. To obtain a sparse $\mathbf{C}^{-1}$ one needs to approximate $\mathbf{C}$ with a diagonal matrix $\widetilde{\mathbf{C}} = \langle \psi_j, 1 \rangle$ with elements $\widetilde{\mathbf{C}}_{jj} = \int \psi_j(\mathbf{s})\mathrm{d}\mathbf{s} = \sum_{i=1}^{r} \mathbf{C}_{ji}$. The resulting precision matrix $\mathbf{Q} = \frac{1}{\tau^2}\mathbf{K}^T\widetilde{\mathbf{C}}^{-1}\mathbf{K}$ is now sparse since $\mathbf{G}$ is sparse ($\mathbf{G}$ contains the finite difference approximation of Δ). For $\alpha = 2$ ($\nu = 1$) the local structure of the precision matrix on a regular grid in $\mathbb{R}^2$ is given by

$$\kappa^4 \underbrace{\begin{bmatrix} & 1 & \\ & & \end{bmatrix}}_{\mathbf{C}} + 2\kappa^2 \underbrace{\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}}_{\approx -\Delta(\text{or } \mathbf{G})} + \underbrace{\begin{bmatrix} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{bmatrix}}_{\approx \Delta^2(\text{or } \mathbf{G_2}=\mathbf{G}\widetilde{\mathbf{C}}^{-1}\mathbf{G})}.$$

Although solving the SPDE might be analytically difficult for complex sets of basis functions, in a comparison study Bolin and Lindgren (2013) showed the SPDE method using simple triangulation basis functions is generally more efficient and accurate compare to covariance tapering and convolution method.

A special case of the GMRF is the intrinsic GMRF, obtained when $\kappa = 0$ in (10), implying an infinite range, this is equivalent to a Wahba smoothing spline (Kimeldorf and Wahba, 1970, Nychka, 2000, Wahba, 1981).

3.4 Inference using INLA and MCMC

To obtain full inference of the posterior distribution of $\mathbf{x}$ in (6), and parameters in (5), Markov Chain Monte Carlo (MCMC) is often used. An alternative to MCMC is to use the Integrated Nested Laplace Approximations (INLA) introduced by Rue et al. (2009). INLA provides faster inference compared to MCMC and is available as open source through the R-INLA package (Lindgren and Rue, 2015). However, R-INLA includes only very limited methods for multivariate non-Gaussian data. For example the Dirichlet distribution used when modelling

compositional data (8) falls outside of the current INLA framework. In this thesis, we use INLA for the model introduced in (7) and MCMC for the model in (8).

INLA is based on Laplace approximations of the data log-likelihood, $f(\mathbf{x}) = \log \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, around the mode, $\mathbf{x}_0$, of the posterior $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$. To obtain an approximation of the posterior in (6), the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ is first approximated. To avoid computing the integral in (5), the expression $\pi(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$ is rewritten as

$$\begin{aligned} \pi(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) &= \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) = \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta}) \\ &\Rightarrow \pi(\mathbf{y}|\boldsymbol{\theta}) = \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})}{\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})}. \end{aligned} \tag{13}$$

For a model with Gaussian data the denominator in (13) has a closed Gaussian form, allowing explicit computation of $\pi(\mathbf{y}|\boldsymbol{\theta})$ for any $\mathbf{x}_0$. For a model with non-Gaussian data there is no closed form expression for the denominator in (13). However, $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ can be approximated by a Gaussian density $\pi_{\mathbf{G}}$ using a Laplace approximation of the data log-likelihood around the mode $\mathbf{x}_0$. The Laplace approximation results in the following approximation of the posterior

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{y}|\mathbf{x}_0, \boldsymbol{\theta})\pi(\mathbf{x}_0|\boldsymbol{\theta})}{\pi_{\mathbf{G}}(\mathbf{x}_0|\mathbf{y}, \boldsymbol{\theta})}\pi(\boldsymbol{\theta}), \qquad \mathbf{x}_0 = \arg\max_{\mathbf{x}} \pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}).$$

Having an approximation of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and using numerical integration over $\boldsymbol{\theta}$, the posterior $\pi(\mathbf{x}_i|\mathbf{y})$ is computed as

$$\pi(\mathbf{x}_i|\mathbf{y}) \propto \int \pi(\mathbf{x}_i|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \approx \sum_j \pi_{\mathbf{G}}(\mathbf{x}_i|\mathbf{y}, \boldsymbol{\theta}_j)\tilde{\pi}(\boldsymbol{\theta}_j|\mathbf{y})$$

where a second Laplace approximation is used for $\pi(\mathbf{x}_i|\mathbf{y}, \boldsymbol{\theta})$ (for details, see Rue et al., 2009). Note that INLA does not compute the full joint posterior $\pi(\mathbf{x}|\mathbf{y})$, only the univariate posteriors $\pi(\mathbf{x}_i|\mathbf{y})$ are computed. To make the Laplace approximation tractable, INLA requires $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_i \pi(\mathbf{y}_i|\mathbf{x}_i, \theta)$, i.e each univariate observation is *only* influenced by *one* point in the latent field. This implies that the Dirichlet data model used for compositional data cannot be approximated using INLA due to the non-linear dependence of each Dirichlet observations on several elements of the latent field.

In MCMC, a Metropolis-Hastings (MH) type algorithm is used (Brooks et al., 2011). Given a specified target distribution with density $p(\mathbf{x})$, the MH updates propose a move to a new state, $\mathbf{x}^*$, from the current state, $\mathbf{x}^o$, with the

proposal probability $q(\mathbf{x}^*|\mathbf{x}^o)$. Then the proposal $\mathbf{x}^*$ is accepted with the acceptance probability $\phi = \min\left(1, \frac{p(\mathbf{x}^*)}{p(\mathbf{x}^o)} \frac{q(\mathbf{x}^o|\mathbf{x}^*)}{q(\mathbf{x}^*|\mathbf{x}^o)}\right)$ or rejected with probability $1-\phi$. A popular proposal for the MH is a symmetric Random-walk (RW) with proposals of the form $\mathbf{x}^* = \mathbf{x}^o + \epsilon$ with $\epsilon \sim h$ (h is often Gaussian, i.e. $\epsilon \sim \mathsf{N}\left(0, \sigma_\epsilon^2\right)$). The proposal probability for a symmetric RW is $q(\mathbf{x}^*|\mathbf{x}^o) = h(\mathbf{x}^* - \mathbf{x}^o)$ and the acceptance probability simplifies to $\phi = \min\left(1, \frac{p(\mathbf{x}^*)}{p(\mathbf{x}^o)}\right)$.

To create fast MCMC inference for the model in (8), the Metropolis Adjusted Langevin algorithm (MALA; Girolami and Calderhead, 2011, Roberts and Stramer, 2003) is used to estimate the process model $\mathbf{x}$,

$$q(\mathbf{x}^*|\mathbf{x}^o) \sim \mathsf{N}\left(\mathbf{x}^o + \frac{\varepsilon^2}{2}\boldsymbol{\mathcal{I}}^{-1}\nabla l, \varepsilon^2\, \boldsymbol{\mathcal{I}}^{-1}\right), \tag{14}$$

where ε is the step size of MALA, ∇l is a vector of derivatives of the log posterior of $\pi(\mathbf{x}|\mathbf{y})$ w.r.t. $\mathbf{x}$ and $\boldsymbol{\mathcal{I}}$ is the expected Fisher information matrix w.r.t the data; i.e the expectation of the second derivatives of the log posterior $\pi(\mathbf{x}|\mathbf{y})$. At each iteration, $\boldsymbol{\mathcal{I}}^{-1}\nabla l$ gives a sampling direction from the current state similar to a Newton-Raphson step (Givens and Hoeting, 2012, Ch. 2), and the proposal variance, $\boldsymbol{\mathcal{I}}^{-1}$, accounts for the dependency among the parameters. Due to the GMRF structure of the latent fields, $\boldsymbol{\mathcal{I}}$ will be a sparse matrix reducing the computational costs.

Example 3.1. *As an example, we sample $N = 30$ observations $\mathbf{x}$ from the normal distribution $\mathsf{N}\left(0, 10^2\right)$. Given the observations we want to estimate μ and σ using a MALA proposal* (14)*. The posterior of μ and σ given $\mathbf{x}$ with flat priors; i.e. $\pi(\mu) = \pi(\sigma) \propto 1$, simplifies to $\mathsf{P}(\mu, \sigma|\mathbf{x}) \propto \prod_{i=1}^N \mathsf{P}(x_i|\mu, \sigma)$. Given the Gaussian distribution, the log posterior, l, and its derivatives, ∇l and $\mathcal{I}$, are obtained as*

$$l = -N\log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^N (x_i-\mu)^2}{2\sigma^2},$$

$$\nabla l_\mu = \frac{\sum_{i=1}^N (x_i-\mu)}{\sigma^2}, \qquad \nabla l_\sigma = \frac{-N}{\sigma} + \frac{\sum_{i=1}^N (x_i-\mu)^2}{\sigma^3},$$

$$\mathcal{I} = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{2N}{\sigma^2} \end{bmatrix}, \qquad \mathcal{I}^{-1} = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{\sigma^2}{2N} \end{bmatrix}.$$

The proposals for the parameters are computed as

$$\mu^* = \mu^o + \frac{\varepsilon^2}{2}\left(\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}\right) + \varepsilon\left(\frac{\sigma}{\sqrt{N}}\right)v, \qquad v \sim \mathsf{N}(0,1),$$

$$\sigma^* = \sigma^o + \frac{\varepsilon^2}{2}\left(\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2N\sigma} - \frac{\sigma}{2}\right) + \varepsilon\left(\frac{\sigma}{\sqrt{2N}}\right)u, \quad u \sim \mathsf{N}(0,1).$$

Figure 3 shows the results of 200 iterations with starting point, $\mu_0 = 5$ *and* $\sigma_0 = 15$ *of a MALA with* $\varepsilon = 0.75$ *and acceptance rate* 55%*, compared to a RW proposal for* μ *and* σ *with* σ_ϵ *being 1 and* 0.1 *respectively, and acceptance rate* 43%*.*

In order to get reasonable acceptance rate, an adaptive MCMC method (Andrieu and Thoms, 2008) is used for the step size of the proposals, with the following updating rule;

$$\varepsilon_{i+1} = \varepsilon_i + \gamma_{i+1}(\widehat{\phi}_{\mathbf{x}}(\varepsilon_i) - \phi)$$

where ε_i is the step size for the i^{th} MCMC iteration, $\gamma_i = i^{-t}$ for $t \in (0,1)$, $\widehat{\phi}_{\mathbf{x}}$ is the acceptance probability of the i^{th} step, and ϕ is the target acceptance rate. The step sizes of MALA proposal and random walk are adjusted to maintain 57% and 40% acceptance rate, respectively (Roberts et al., 2001).

3.5 Posterior uncertainties for the compositions

To compute the uncertainty in the posterior of $\mathbf{x}$ and therefore in the compositional reconstructions, a novel way of constructing joint confidence regions for the entire composition at each location is proposed (Pirzamanbein et al., 2015). Using the MCMC samples of $\mathbf{x}$, along with the mean, $\boldsymbol{\mu}_{\mathbf{x}}$, and the covariance, $\boldsymbol{\Sigma}_{\mathbf{x}}$, of the samples, we construct the elliptical confidence region (CR) for a multivariate Gaussian distribution as

$$(\mathbf{x} - \mu_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}) = M_\alpha.$$

For a multivariate Gaussian field, M_α is taken as a suitable quantile of a chi-squared distribution, $\chi_\alpha(d)$. Since the posterior of $\mathbf{x}|\mathbf{y}$ is approximately Gaussian, M_α is chosen as the α sample quantile of the above squared Mahalanobis distance computed for all MCMC samples of $\mathbf{x}|\mathbf{y}$. Given a confidence region in $\mathbb{R}^2$, the inverse transform function (2) is used to transfer CR to ternary confidence

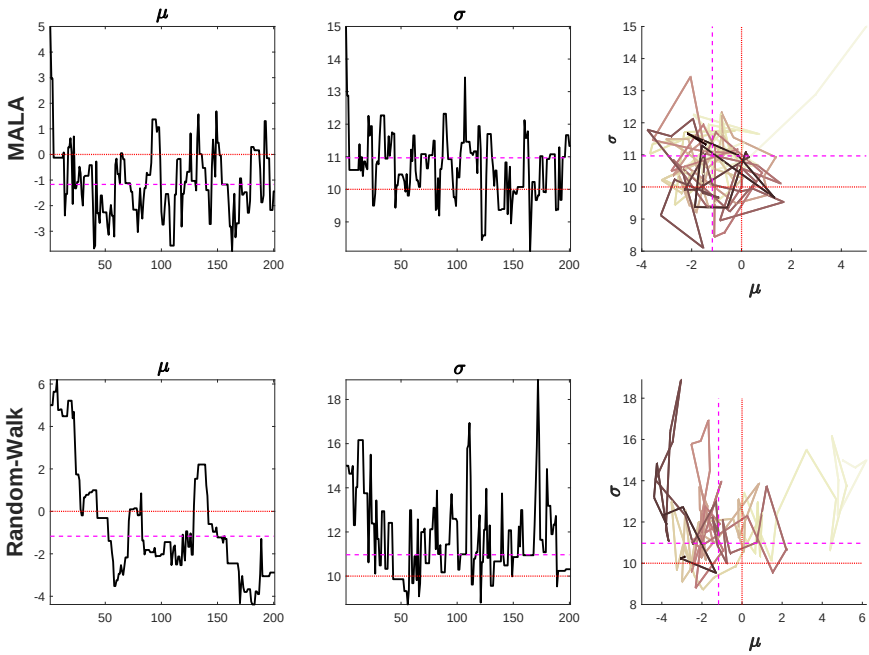


Figure 3: Estimated μ and σ given 30 observations of a Gaussian distribution based on row 1) MALA proposal and row 2) random walk proposal. The dashed line indicates the maximum likelihood estimation of the parameters and the dotted line shows the true parameter values used for simulating the observations. The color shading in the last column shows the evolution of the chain from light to dark color.

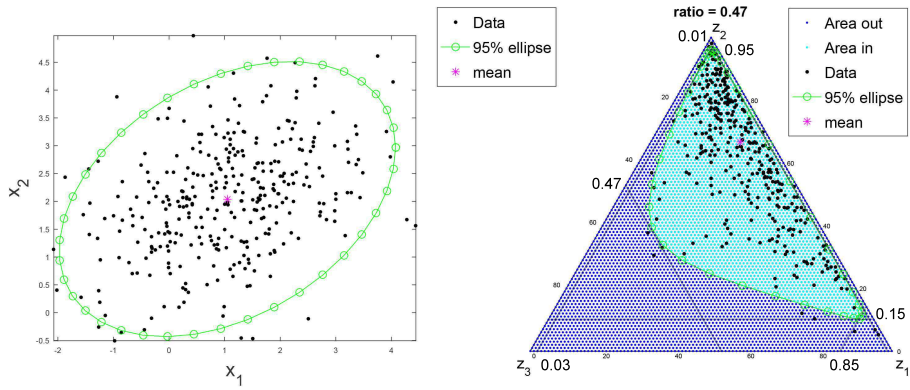


Figure 4: The left plot shows the 95% elliptical confidence region. The right ternary diagram shows the transformed ellipse together with the ratio compared to unit triangle and maximum/minimum along each component.

region in $(0,1)^3$. In this way, the correlation between the components of the multivariate field is taken into account. Figure 4 shows the 95% CR for simulated 3-compositional data. In order to comprehend the size of the uncertainty to evaluate the performance of the model, the ratio of the ternary CR to the entire triangle is computed by distributing points in the ternary diagram and calculating the proportion of the points inside the CR to the unit triangle. To illustrate the changes in compositions, the maximum and minimum along each dimensions of the ternary plot, i.e. in each component is used. In this way, a joint lower bound (minimum) and upper bound (maximum) for each composition together with the corresponding, most likely changes in the other compositions is obtained.

In addition, prediction region (PR) for the compositions is also computed to evaluate the predictive performance of the model. To obtain PR, new Dirichlet observations are simulated for each MCMC sample of $\mathbf{x}|\mathbf{y}$ and $\boldsymbol{\alpha}|\mathbf{y}$. These D-compositional Dirichlet-simulations are then transformed to $\mathbb{R}^2$ using the transform function (1). The procedure above can then be used to obtain prediction ellipses and corresponding ternary PRs.

4 Outline of the papers

Paper A

Creating spatially continuous maps of past land cover from point estimates: A new statistical approach applied to pollen data

Behnaz Pirzamanbein, Johan Lindström, Anneli Poska, Shinya Sugita, Anna-Kari Trondman, Ralph Fyfe, Florence Mazier, Anne B. Nielsen, Jed O. Kaplan, Anne E. Bjune, H. John B. Birks, Thomas Giesecke, Mikhel Kangur, Małgorzata Latałowa, Laurent Marquer, Benjamin Smith and Marie-José Gaillard

In Paper A, the pollen based land cover data are reconstructed using the model introduced in (7) for three time periods. In order to assess the importance of the spatial dependence structure in the data, two versions of the model are used; 1) an intrinsic GMRF with $\kappa = 0$ and separable covariance structure $\boldsymbol{\rho} \otimes \mathbf{Q}^{-1}$ where $\boldsymbol{\rho}$ is a matrix of covariances among the multivariate fields $\mathbf{x}$, and $\mathbf{Q}$ is a precision matrix of a GMRF

$$\begin{aligned} \mathbf{u}(s_i)|\mathbf{x}, \sigma_\epsilon &= \mathsf{N}\left(\mathbf{x}(s_i), \sigma_\epsilon^2\right) \qquad i = 1, \cdots N \\ \mathbf{x}|\boldsymbol{\rho}, \boldsymbol{\beta} &\sim \mathsf{N}\left(\mathbf{B}\boldsymbol{\beta}, \boldsymbol{\rho} \otimes \mathbf{Q}^{-1}\right) \end{aligned}$$

and 2) a regression model consisting of only the regression part, $\mathbf{x} = \mathbf{B}\boldsymbol{\beta}$, without any spatial dependence structure. For both models, two different sets of possible covariates are evaluated. Both sets of covariates include intercept, elevation and a combination of vegetation estimates from a DVM (LPJ-GUESS; Smith et al., 2001) and an ALCC estimates (KK10; Kaplan et al., 2009). The second covariate set also included geographical coordinates (longitude and latitude).

The parameters of the model and the latent process $\mathbf{x}$ are estimated using INLA (Rue et al., 2009) as implemented by the R-INLA package (Lindgren and Rue, 2015) in R (R Core Team, 2014).

The performance of the model was evaluated by computing ACD, using (3), between a present time forest map of Europe and the model reconstructions for the 1900 CE time period. The model was also evaluated using 6-fold cross-validation for all time periods.

The results indicated that the IGMRF models perform better than the regression models and that the smaller covariates set provides the best reconstructions of past land cover compositions.

Regarding the theory and implementation, the model was developed in collaboration between Behnaz Pirzamanbein and Johan Lindström. Behnaz Pir-

zamanbein implemented the methods. The writing was done mainly by Behnaz Pirzamanbein, Johan Lindström, Anneli Poska, and Marie-José Gaillard. Anneli Poska, and Marie-José Gaillard helped with the introduction and conclusions regarding the environmental and biological interpretation of data and results. In addition, Shinya Sugita helped with the structure of the paper and text clarification. The rest of the co-authors contributed with the data collection, data preparation and proofreading the text.

Paper B

Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties

Behnaz Pirzamanbein, Johan Lindström, Anneli Poska, and Marie-José Gaillard

In Paper B, the compositional data are modelled using the model described in (8). Using the Dirchlet distribution, the model aims to correctly estimate the uncertainties in the composition, which might be underestimated using the model in Paper A. Similar to Paper A, two different models are used for modelling the land cover compositions;1) Full model

$$\begin{aligned}
&\mathbf{y}(s_i)|\mathbf{x}, \alpha \sim \mathsf{Dir}\left(\alpha, f(\mathbf{x}(s_i))\right) \qquad i = 1, \cdots N \\
&\mathbf{x}|\beta, \kappa, \rho \sim \mathsf{N}\left(\mathbf{B}\beta, \rho \otimes \mathbf{Q}^{-1}(\kappa)\right)
\end{aligned}$$

and 2) regression model consisting of only mean structure with out spatial dependency. For the mean structure, the best covariates set from Paper A is used which includes intercept, elevation and a combination of vegetation estimates from a DVM (LPJ-GUESS; Smith et al., 2001) and an ALCC estimates (KK10; Kaplan et al., 2009).

In contrast with Paper A the inference is not possible using R-INLA (Lindgren and Rue, 2015), due to the non-linear dependency of the Dirchlet observations on several latent fields. Therefore, a MCMC algorithm is used to estimate all the parameters and the latent fields. In MCMC, the parameters are divided into two blocks; in the first block α and $\mathbf{x}$ are updated using MALA proposal with adaptive step size and in the second block, parameters of latent field, κ and ρ are updated.

The maps of past land cover for five time periods are reconstructed. A novel way of computing the uncertainties for composition is proposed (Section 3.5) and the confidence and predictive regions are computed. The model evaluation is done using a 6-fold cross validation by computing the ACD (3) for all the time

periods. The reconstruction for the 1900 CE is also compared to the recent time periods European forest map and also with the reconstructions based on Gaussian model explained in Paper A.

The results show that the Dirichlet observation model with spatial dependency developed in this paper performs best in reconstructing the land cover compositions.

Behnaz Pirzamanbein developed and implemented the methods and also wrote most of the paper. Contributions from the other authors were as follows. Regarding the theory and implementation, Johan Lindström helped with details regarding the joint uncertainties as well as suggestions for debugging, including common implementation mistakes. Johan also helped with polishing of the text and wrote parts of the introduction and conclusion. The two other co-authors, Anneli Poska and Marie-José Gaillard, are from "Physical Geography and Ecosystems Analysis" and "Biology, Palaeoecology and Environmental sciences", respectively. They provided the pollen data and vegetation model output. They also helped with the introduction and conclusions regarding the environmental and biological interpretation of the data and results.

Paper C

Analysing the sensitivity of pollen based land cover maps to different auxiliary variables

Behnaz Pirzamanbein, Anneli Poska, Johan Lindström

One of the covariates used in the model developed in Paper B is a combination of vegetation estimates from the DVM LPJ-GUESS and the ALCC estimates of KK10. Due to differences in climate forcing used in the DVM and significant variation in land use estimates between the existing ALCC, this covariate can vary substantially. In Paper C, the sensitivity and robustness of the model to the choice of covariates is analysed. Different covariates including elevation, and different combinations of two different sets of ALCC scenarios, KK10 and HYDE, and two different estimations of natural vegetation from the DVM LPJ-GUESS based on two climate forcing, RCA3 (Rossby Centre Regional Climate Model Samuelsson et al., 2011) and Earth System Model (ESM Mikolajewicz et al., 2007) are used.

For evaluating the model results, the ACDs (3) are computed among the models reconstructions. For the 1900 CE time period, the ACDs are computed between the land cover reconstructions from each model and the present day European forest map. Since there is no ground truth data available for the 1725

CE and 4000 BCE time periods, a 6-fold cross validation is used by computing ACDs. To measure the predictive performances of the models, deviance information criteria (DIC) is also computed for each model and time period.

The results show that although there are small differences in the land cover reconstructions, the overall performance of the models are not sensitive to the choice of covariates. All the tests indicated the similarity between the model reconstructions. This is due to the fact that the changes in covariate can be corrected by different regression coefficient estimates and even though being different the model based covariates share similar spatial patterns which help the reconstruction.

Behnaz Pirzamanbein set up the analysis in collaboration with the co-authors. The implementations, results and the statistical analysis were provided by Behnaz Pirzamanbein. Moreover, she wrote the method section, some parts of introduction and conclusion. Anneli Poska provided the data. She also wrote most of the introduction and conclusions regarding the environmental and biological interpretation of the data and results. Johan Lindström helped with structure, polishing the text, and parts of the introduction and conclusion.

Paper D

Reconstruct past Human Land Use from Pollen data and Anthropogenic Land Cover Changes scenarios

Behnaz Pirzamanbein, Johan Lindström

In Paper D, the possibility of combining the pollen based land cover composition (LCC) with population based estimates of anthropogenic land cover changes (ALCC) was investigated by extending the model in Paper B. The aim of the paper was to merge LCC and ALCC to provide estimates of past natural land cover and past human land use.

The available ALCC estimates consist of 1) KK10 (Kaplan et al., 2009), and 2) HYDE (Klein Goldewijk et al., 2011). Since KK10 and HYDE are substantially different in older time periods, perturbed proportions of human land used based on both datasets were added to the data model in the hierarchy. The full

hierarchical model is formulated as

$$\mathbf{L}|\alpha, \mathbf{z} \sim \mathsf{Dir}\,(\alpha, \mathbf{z})\,, \qquad \mathbf{H}_k|\lambda, \mathbf{p}_{H,k} \sim \mathsf{Beta}\left(\lambda \mathrm{p}_{H,k}, \lambda(1-\mathrm{p}_{H,k})\right),$$

$$\mathbf{z} = h(\mathbf{p}_L, \mathrm{p}_H),$$

$$\mathbf{p}_L = f(\mathbf{x}_L),$$
$$\mathrm{p}_H = g(\mathrm{x}_H), \qquad \mathrm{p}_{H,k} = g(\mathrm{x}_H + \epsilon_k),$$

$$\mathbf{x}|\kappa, \boldsymbol{\rho} \sim \mathsf{N}\left(\mathbf{B}\boldsymbol{\beta}, \boldsymbol{\rho} \otimes \mathbf{Q}^{-1}(\kappa)\right).$$

Two versions of covariates are used in the modelling, one included the LPJ-GUESS natural vegetation as covariate for natural land cover compositions and the other one only included elevation. In the later case the natural land cover reconstruction can be compared with the LPJ-GUESS natural vegetation.

The model is fitted using a block updated MCMC. In the first block the concentration parameters of beta and Dirichlet distribution, α and λ, together with latent process $\mathbf{x}$ and ϵ_k were updated using a MALA proposal. In the second block the parameters of latent fields including κ and $\boldsymbol{\rho}$ were updated using RW and conjugacy. Finally, the precision of added perturbation, τ, was updated using conjugacy.

The model results are evaluated using leave one block out. The results showed that it is possible to combine pollen based land cover and ALCC scenarios to obtain past human land use and natural land cover. Further, the results indicate the human land use reconstructions based on both ALCC and pollen based LCC are closer to the HYDE proportions than KK10, in contrast the spatial pattern is closer to KK10 than HYDE. This suggests that pollen based LCC can be used to adjust the existing population based ALCC estimates to match observed past vegetation patterns and recover past human land use from pollen based LCC.

Initially, a model based on archaeological artefacts instead of ALCC was attempted to reconstruct the human land use using log Gaussian cox process (Simpson et al., 2016). Unfortunately due to biases in our archaeological data and issues with their dating, the model could not fit to the actual data. However, the preliminary results indicated modelling the human land use using archaeological artefacts is possible and improvement in the archaeological data sets might give better reconstructions of past human land use compare to the existing ones.

The problem formulation was done in collaboration with Johan Lindström. Behnaz Pirzamanbein implemented most of the methods and wrote different parts of the text. Johan Lindström helped with different part of the text. He also helped with debugging the implementation.

References

J. Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., 1986.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. and Comput.*, 18(4):343–373, 2008.

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *J. Roy. Statist. Soc. Ser. B*, 70(4):825–848, 2008.

D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *J. Am. Statist. Assoc.*, 96(456):1205–1214, 2001.

D. Bolin and F. Lindgren. A comparison between Markov approximations and other methods for large spatial data sets. *Comput. Statist. and Data Anal.*, 61: 7–21, 2013.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

M. Claussen, V. Brovkin, and A. Ganopolski. Biogeophysical versus biogeochemical feedbacks of large-scale land cover change. *Geophys. Res. Lett.*, 28(6):1011–1014, 2001.

N. Cressie and G. Johannesson. Fixed rank Kriging for very large spatial data sets. *J. Roy. Statist. Soc. Ser. B*, 70(1):209–226, 2008.

R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.*, 15(3):502–523, 2006.

M.-J. Gaillard, S. Sugita, F. Mazier, A.-K. Trondman, A. Brostrom, T. Hickler, J. O. Kaplan, E. Kjellström, U. Kokfelt, P. Kuneš, , C. Lemmen, P. Miller, J. Olofsson, A. Poska, M. Rundgren, B. Smith, G. Strandberg, R. Fyfe, A. Nielsen, T. Alenius, L. Balakauskas, L. Barnekov, H. Birks, A. Bjune, L. Björkman, T. Giesecke, K. Hjelle, L. Kalnina, M. Kangur, W. van der Knaap, T. Koff, P. Lagerås, M. Latałowa, M. Leydet, J. Lechterbeck, M. Lindbladh, B. Odgaard, S. Peglar, U. Segerström, H. von Stedingk, and H. Seppä. Holocene land-cover reconstructions for studies on land cover-climate feedbacks. *Clim. Past.*, 6:483–499, 2010.

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 73(2):123–214, 2011.

G. H. Givens and J. A. Hoeting. *Computational statistics*, volume 710. John Wiley & Sons, 2012.

S. Hellman, M.-j. Gaillard, A. Broström, and S. Sugita. Effects of the sampling design and selection of parameter values on pollen-based quantitative reconstructions of regional vegetation: a case study in southern Sweden using the REVEALS model. *Veg. Hist. Archaeobot.*, 17(5):445–459, 2008.

D. Higdon. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. and Ecological Statist.*, 5(2):173–190, 1998.

J. O. Kaplan, K. M. Krumhardt, and N. Zimmermann. The prehistoric and preindustrial deforestation of Europe. *Quaternary. Sci. Rev.*, 28(27):3016–3034, 2009.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2): 495–502, 1970.

K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Global. Ecol. Biogeogr.*, 20(1):73–86, 2011.

F. Lindgren and H. Rue. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.*, 63(19):1–25, 2015. doi: 10.18637/jss.v063.i19.

F. Lindgren, R. Håvard, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc. Ser. B*, 73(4):423–498, 2011.

B. Matérn. Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelanden från statens Skogsforskningsinstitut 49, Statens Skogsforsningsinstitut, Stockholm, Sweden, 1960.

U. Mikolajewicz, M. Gröger, E. Maier-Reimer, G. Schurgers, M. Vizcaíno, and A. M. Winguth. Long-term effects of anthropogenic co2 emissions simulated with a complex earth system model. *Clim. Dynam.*, 28(6):599–633, 2007.

D. W. Nychka. Spatial-process estimates as smoothers. In M. G. A. Schimek, editor, *Smoothing and Regression: Approaches, Computation, and Application*, pages 393–424. Wiley, New York, USA, 2000.

C. J. Paciorek and J. S. McLachlan. Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Am. Statist. Assoc.*, 104(486):608–622, 2009.

B. Pirzamanbein, J. Lindström, A. Poska, and M.-J. Gaillard. Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *arXiv preprint arXiv:1511.06417*, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

G. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4(4):337–358, 2003.

G. O. Roberts, J. S. Rosenthal, et al. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Statist. Soc. Ser. B*, 71(2):319–392, 2009.

P. Samuelsson, C. G. Jones, U. Willén, A. Ullerstig, S. Gollvik, U. Hansson, C. Jansson, E. Kjellström, G. Nikulin, and K. Wyser. The rossby centre regional climate model rca3: model description and performance. *Tellus A*, 63 (1):4–23, 2011.

D. Simpson, J. Illian, F. Lindgren, S. Sorbye, and H. Rue. Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103 (1):49–70, 2016.

B. Smith, I. C. Prentice, and M. T. Sykes. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Global. Ecol. Biogeogr.*, 10(6):621–637, 2001.

G. Strandberg, E. Kjellström, A. Poska, S. Wagner, M.-J. Gaillard, A.-K. Trondman, A. Mauri, B. A. S. Davis, J. O. Kaplan, H. J. B. Birks, A. E. Bjune, R. Fyfe, T. Giesecke, L. Kalnina, M. Kangur, W. O. van der Knaap, U. Kokfelt, P. Kuneš, M. Latałowa, L. Marquer, F. Mazier, A. B. Nielsen, B. Smith, H. Seppä, and S. Sugita. Regional climate model simulations for Europe at 6 and 0.2 k bp: sensitivity to changes in anthropogenic deforestation. *Clim. Past.*, 10(2):661–680, 2014.

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

H. Tjelmeland and K. V. Lund. Bayesian modelling of spatial compositional data. *J. Appl. Stat.*, 30(1):87–100, 2003.

A.-K. Trondman, M.-J. Gaillard, F. Mazier, S. Sugita, R. Fyfe, A. B. Nielsen, C. Twiddle, P. Barratt, H. J. B. Birks, A. E. Bjune, L. Björkman, A. Broström, C. Caseldine, R. David, J. Dodson, W. Dörfler, E. Fischer, B. van Geel, T. Giesecke, T. Hultberg, L. Kalnina, M. Kangur, P. van der Knaap, T. Koff, P. Kuneš, P. Lagerås, M. Latałowa, J. Lechterbeck, C. Leroyer, M. Leydet, M. Lindbladh, L. Marquer, F. J. G. Mitchell, B. V. Odgaard, S. M. Peglar, T. Persson, A. Poska, M. Rösch, H. Seppä, S. Veski, and L. Wick. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. *Glob. Change Biol.*, 21(2):676–697, 2015.

G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Comput.*, 2:5–16, 1981.

P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

P. Whittle. Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):975–994, 1963.

C. K. Wikle. Hierarchical modeling with Spatial Data. In *Handbook of Spatial Statistics*, pages 93–111. CRC press, 2011.

A

Paper A

Creating Spatially Continuous Caps of Past Land Cover from Point Estimates: A New Statistical Approach Applied to Pollen Data

Behnaz Pirzamanbein[1], Johan Lindström[1], Anneli Poska[1,2], Shinya Sugita[3], Anna-Kari Trondman[4], Ralph Fyfe[5], Florence Mazier[6], Anne B. Nielsen[1], Jed O. Kaplan[7], Anne E. Bjune[8], H. John B. Birks[9,10,11], Thomas Giesecke[12], Mikhel Kangur[3], Małgorzata Latałowa[13], Laurent Marquer[4], Benjamin Smith[1], Marie-José Gaillard[4]

[1] *Lund University, Sweden* [2] *Tallinn University of Technology, Estonia* [3] *Tallinn University, Estonia* [4] *Linnaeus University, Sweden* [5] *University of Plymouth, UK* [6] *University of Toulouse, France* [7] *University of Geneva, Switzerland* [8] *Uni Climate, Uni Research AS and Bjerknes Centre for Climate Research, Norway* [9] *University of Bergen, Norway* [10] *University College London, UK* [11] *University of Oxford, UK* [12] *University of Göttingen, Germany* [13] *University of Gdańsk, Poland*

Abstract

Reliable estimates of past land cover are critical for assessing potential effects of anthropogenic land-cover changes on past earth surface-climate feedbacks and landscape complexity. Fossil pollen records from lakes and bogs have provided important information on past natural and human-induced vegetation cover. However, those records provide only point estimates of past land cover, and not the spatially continuous maps at regional and sub-continental scales needed for climate modelling.

We propose a set of statistical models that create spatially continuous maps of past land cover by combining two data sets: 1) pollen-based point estimates of past land cover (from the REVEALS model), and 2) spatially continuous estimates of past land cover, obtained by combining simulated potential vegetation (from LPJ-GUESS) with an anthropogenic land-cover change scenario (KK10). The proposed models rely on statistical methodology for compositional data and use Gaussian Markov Random Fields to model spatial dependencies in the data.

Land-cover reconstructions are presented for three time windows in Europe: 0.05 ka, 0.2 ka, and 6 ka years before present (BP). The models are evaluated through cross-validation, deviance information criteria and by comparing the reconstruction of the 0.05 ka time window to the present-day land-cover data compiled by the European Forest Institute (EFI). For 0.05 ka, the proposed models provide reconstructions that are closer to the EFI data than either the REVEALS- or LPJ-GUESS/KK10-based estimates; thus the statistical combination of the two estimates improves the reconstruction. The reconstruction by the proposed models for 0.2 ka is also good. For 6 ka, however, the large differences between the REVEALS- and LPJ-GUESS/KK10-based estimates reduce the reliability of the proposed models. Possible reasons for the increased differences between REVEALS and LPJ-GUESS/KK10 for older time periods and further improvement of the proposed models are discussed.

Key words: Land cover, Spatial modeling, Paleoecology, Pollen, Compositional data, Gaussian Markov random fields

1 Introduction

Anthropogenic impacts on past land cover have potentially influenced the climate system more significantly than previously assumed (e.g. Ruddiman, 2005). Many simulation studies have evaluated the biogeophysical effects of vegetation and land-use changes on past climate at the global scale (e.g. Brovkin et al., 2006, Christidis et al., 2013, Claussen et al., 2001, de Noblet-Ducoudré et al., 2012, Pitman et al., 2009, Pongratz et al., 2010). However, descriptions of past land cover vary considerably among studies, including: static present-day land cover (Strandberg et al., 2011), dynamic (or static) potential land cover simulated by dynamic vegetation models (DVMs) (e.g. Brovkin et al., 2002, Hickler et al., 2012), and land-cover estimates combining DVMs and anthropogenic land-cover change (ALCC) scenarios (de Noblet-Ducoudré et al., 2012, Pongratz et al., 2009). The

existing ALCC scenarios (e.g. Kaplan et al., 2009, Klein Goldewijk et al., 2011, Pongratz et al., 2009) also differ significantly from each other (Gaillard et al., 2010) and their reliability still needs to be evaluated.

Palaeoecology has provided important information on past vegetation and land cover using fossil pollen and plant macroremains deposited and preserved in lake and bog sediments over thousands of years. Although those palaeorecords provide insights into the past vegetation that modelling approaches cannot, the interpretation of palaeorecords, particularly quantification of land-cover changes in specific spatiotemporal scales, remains difficult. In addition palaeorecords are point estimates of land cover around study sites. Therefore, the gaps between estimates at study points need to be filled if palaeorecords of land-cover changes are to be useful in climate modelling and other simulation studies that require quantitative and spatially continuous input datasets. To achieve this interpolation process we propose a new statistical approach based on statistical spatial models and methods developed in Tjelmeland and Lund (2003), Lindgren et al. (2011) and Rue et al. (2009). Our approach takes spatially continuous estimates of past land cover from a DVM and an ALCC scenario as covariates and then constrains those using the point estimates of pollen-based land cover; thus it can potentially avoid problems that conventional interpolation methods using fossil pollen records have. The DVMs and ALCC scenarios provide a way of capturing land-cover changes due to the non-stationary environmental conditions in Europe over areas with few or no pollen-based observations.

This paper aims at reconstructing the land cover in Europe at 6.0 ka, 0.2 ka and 0.05 ka (calibrated year BP) using the methods developed in this study. The work is part of the LANDCLIM project (LAND cover - CLIMate interactions in Europe during the Holocene; Gaillard et al., 2010) that assesses the possible effects of long-term changes in anthropogenic land cover on the Holocene climate (Strandberg et al., 2014). Our objective is also to provide methods and reconstructions that can be used in the evaluation of ecological complexity of European landscapes in the past, i.e. give us new insights on the respective roles played by climate, soils, geography, geology and human impact in landscape dynamics at the spatio-temporal resolutions we are working with. Here is a brief roadmap of this paper to explain and help sort out the complex web of different models and datasets used in the analysis:

Section 2 describes a statistical approach for compositional data (Aitchison, 1986) such as land-cover estimates in proportion. To avoid the time-consuming

inference in Tjelmeland and Lund (2003), the spatial dependence is modelled using a Gaussian Markov Random Field (GMRF) (Lindgren et al., 2011) with fast inference obtained through R-INLA (Lindgren and Rue, 2013, Rue et al., 2009). Two standard linear regression models and two GMRF-based models are developed to explain REVEALS land-cover by various sets of covariates (i.e. estimates from a DVM and an ALCC scenario, elevation, longitude and latitude).

Section 3 describes models and databases used for reconstruction of past and recent (0.05 ka) land cover with the new statistical approach and for data-model comparison. Pollen-based estimates of three land-cover types (coniferous, broadleaved and unforested) at $1^\circ \times 1^\circ$ resolution are obtained using the REVEALS model (Sugita, 2007); hereafter those estimates are referred to as grid-based REVEALS (GB-REVEALS). Potential natural vegetation is simulated by a process-based dynamic ecosystem model LPJ-GUESS (Smith et al., 2001), and anthropogenic land cover is extracted from the ALCC KK10 scenario of Kaplan et al. (2009) based on human population history and technology development. KK10 is the existing ALCC scenario that is closest to the pollen-based GB-REVEALS in terms of degree of past deforestation (Kaplan et al., 2014, Strandberg et al., 2014, Trondman et al., 2012). Combined estimates of model-based potential vegetation and ALCC, hereafter referred to as LPJ-GUESS$_{\mathrm{KK10}}$, are used as one of the main covariates in the data analysis. In addition, the present-day land cover is obtained from the land-cover database of the European Forest Institute (EFI).

Section 4 describes the results and section 5 discusses the significance and implications of the approach developed in this study. The reconstruction of recent land cover is compared to the EFI forest map for evaluation, and pros and cons of the new statistical approach are assessed in detail.

2 Development of the Statistical Model

2.1 Methods for compositional data

In each grid cell three land-cover types (LCTs) - coniferous forest, broadleaved forest, and unforested land - are expressed as proportions. To account for the restrictions inherent to compositional data we apply logratio transformation (Aitchison, 1986) for the LCT data.

Letting $y_i(\mathbf{s})$ denote the fraction of the i^{th} LCT at grid cell location $s \in R^2$;

the values have to sum to one and be non-negative, i.e.

$$\sum_{i=1}^{D} y_i(\mathbf{s}) = 1, \quad \text{and} \quad 0 \leq y_i(\mathbf{s}) \leq 1, \ \forall i. \tag{1}$$

These conditions complicate any statistical analysis. A common solution (Aitchison, 1986, Tjelmeland and Lund, 2003) is to transform the data, allowing modelling to proceed without being encumbered by the restrictions in (1). Several possible transformations exist. Here we use the additive logratio (alr) following Tjelmeland and Lund (2003);

$$u_i(\mathbf{s}) = \log \frac{y_i(\mathbf{s})}{y_D(\mathbf{s})}, \quad i = 1, \ldots, D-1, \tag{2}$$

with D denoting the number of components ($D = 3$ for our three LCTs). The alr takes the set of D compositional values in $[0, 1]$ and transforms them into $D-1$ real valued (i.e. unrestricted) data, $u_i(\mathbf{s})$. The original fractions can be recovered from $u_i(\mathbf{s})$ through the inverse transformation:

$$\begin{aligned} y_i(\mathbf{s}) &= \frac{\exp(u_i(\mathbf{s}))}{1 + \sum_i^{D-1} \exp(u_i(\mathbf{s}))}, \quad i = 1, \ldots, D-1, \\ y_D(\mathbf{s}) &= \frac{1}{1 + \sum_i^{D-1} \exp(u_i(\mathbf{s}))}, \end{aligned} \tag{3}$$

where it is easy to see that the $y_i(\mathbf{s})$ obeys the restrictions in (1).

The alr transformation has its own limitations. It requires proportions to be $y_i(\mathbf{s}) > 0$ and $y_i(\mathbf{s}) < 1$ eliminating the possibility of an equality in (1). This limitation is not an issue for the data used in this paper, since all LCTs are present in all grid cells. Note that increasing y_1 implies a lowering of y_2 and y_3 through the sum to one constraint; thus u_1 and u_2 are dependent, an important fact for the modelling that is further discussed in Section 2.2.

To compute the difference between two compositions we use the compositional distance (Aitchison, 1986):

$$\Delta(\mathbf{u}, \mathbf{v}) = [(\mathbf{u} - \mathbf{v})^T \mathbf{H}^{-1} (\mathbf{u} - \mathbf{v})]^{1/2} \tag{4}$$

where u and v are alr transforms of compositions and $\mathbf{H}$ is a $d \times d$-matrix ($d = D - 1$) with elements $h_{ij} = 2$ if $i = j$, and $h_{ij} = 1$ if $i \neq j$.

A convenient way of illustrating the variability of D-compositional data is a ternary diagram (see Aitchison, 1986, ch. 1.4). Figure 1 illustrates the concept using four compositional data points, each containing coniferous forest (C), broadleaved forest (B), and unforested land (U). In a ternary diagram, a point close to a vertex (e.g. $y(3)$ close to U) has large proportion of the corresponding vertex and a point close to each edge (e.g. $y(4)$ close to U-C) has a low proportion of the opposite vertex (B).

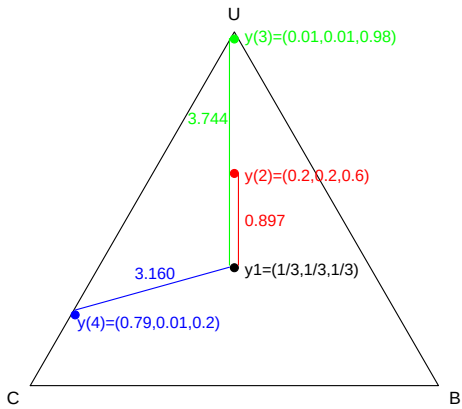


Figure 1: A ternary diagram containing four 3-compositional data points, each consisting of coniferous forest (C), broadleaved forest (B), and unforested land (U). The points correspond to compositions of $y(s) = (y_C(s), y_B(s), y_U(s))$. Numbers along the lines between points indicate their distances according to (4) from $y(1)$. The figure is inspired by Fig. 1. in Billheimer et al. (2001).

2.2 Statistical model

The transformed data, $(u_1, u_2) = \text{alr}(y_C, y_B, y_U)$, is modelled as a multivariate Gaussian process (see Tjelmeland and Lund, 2003),

$$\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix} + \mathbf{A}\begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix} \in \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{z}}), \quad \boldsymbol{\varepsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2). \quad (5)$$

Here $\mathbf{u}_1$ and $\mathbf{u}_2$ are column vectors containing the alr transformed ((2)) compositional values for the N_{obs} observed locations; $\boldsymbol{\mu}_i$ are column vectors of mean values for each location; $\boldsymbol{\varepsilon}$ are independent Gaussian residuals with variance σ_ε^2; and $\mathbf{z}_i$

are spatially dependent residual fields modelling any remaining dependence in the observations, the $N_{\text{obs}} \times N$ matrix $\mathbf{A}$ is a sparse matrix that extracts elements corresponding to the observed locations from $\mathbf{z}_i$. The multivariate Normal model contains two main components: a mean field, $\boldsymbol{\mu}$, and a spatially dependent residual field, $\mathbf{z}$. Those two components are described below.

2.2.1 Mean field

The mean field is modelled as a linear combination of covariates

$$\boldsymbol{\mu}_i = \mathbf{1}\beta_{0,i} + \sum_p \mathbf{B}_p(s)\beta_{p,i} \tag{6}$$

where $\mathbf{B}_p$ is a column vector containing the p^{th} covariate, $\beta_{p,i}$ are unknown regression coefficients and $\mathbf{1}$ is a column vector of ones (the intercept). Two sets of different covariates are used for vegetation reconstruction in this study: $\mathbf{B}$ — contains the alr transformation of $\text{LPJ-GUESS}_{\text{KK10}}$ (LPJ-GUESS estimates adjusted for human impact with the KK10 scenario; see next section) and $\mathbf{B}_{\mathbf{geo}}$ — contains $\text{LPJ-GUESS}_{\text{KK10}}$ and the geographical coordinates, i.e. longitude and latitude. Both $\mathbf{B}$ and $\mathbf{B}_{\mathbf{geo}}$ also take elevation as a covariate. The geographical covariates are fixed over the different time windows, thus possibly adjusting reconstructions for geographically consistent biases in the potential vegetation.

As an alternative to potential vegetation, some bioclimatic covariates (i.e. temperature, precipitation, and soil suitability), used as drivers for LPJ-GUESS, were also included directly in the mean field. Those alternatives did not improve the reconstructions and, for brevity, the results are neither shown nor discussed in this paper.

Using only the mean field, (6) and *without* spatially dependent residual fields (i.e. $\mathbf{z}_i = 0$), the full model (5) reduces to a standard linear regression. We construct two Regression Models **RM** and $\mathbf{RM}_{\mathbf{geo}}$ with $\mathbf{B}$ and $\mathbf{B}_{\mathbf{geo}}$, respectively.

2.2.2 Residual field

The inclusion of coordinates in the mean field only implies a linear dependence between the transformed composition in each grid cell and the corresponding coordinates; no other dependence among neighbouring locations is implied. Any remaining spatial structure can be accounted for by imposing a more complex model for the covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{z}}$, of the residual field $\mathbf{z}$ in (5). Tjelmeland

and Lund (2003) used a Gaussian field (GF) specified through the covariance function, and Paciorek and McLachlan (2009) used a thin plate spline to model the spatial structure. Here we replace the GF with a Gaussian Markov Random Field (GMRF) (Lindgren et al., 2011); this has two main benefits:

1. GMRF has computational benefits over the covariance formulations, and
2. it allows the use of standard software (the R-INLA package Lindgren and Rue, 2013, Rue et al., 2009) for inference.

We now briefly present the GMRF model. According to Whittle (1954, 1963) GFs with Matérn covariance

$$\operatorname{cov}(z(\mathbf{0}), z(\mathbf{s})) = \sigma^2 \frac{(\kappa\|\mathbf{s}\|)^\nu K_\nu(\kappa\|\mathbf{s}\|)}{\Gamma(\nu)2^{\nu-1}} \tag{7}$$

are the solutions to Stochastic Partial Differential Equation (SPDE)

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} z(\mathbf{s}) = \tau\mathcal{W}(\mathbf{s}), \qquad \mathbf{s} \in \mathbb{R}^d, \quad \alpha = \nu + d/2, \tag{8}$$

where $\mathcal{W}(\mathbf{s})$ is Gaussian white noise, $\Delta = \dfrac{\partial^2}{\partial s_x^2} + \dfrac{\partial^2}{\partial s_y^2}$ is the Laplacian, κ is the spatial scale parameter, ν controls the smoothness, τ controls the variance of z and is linked to σ^2 (see Lindgren et al., 2011, ch. 2.1), and K_ν is the modified Bessel function of the second kind.

Lindgren et al. (2011) showed that a GMRF representation of a Matérn GF can be explicitly constructed with precision matrix $\mathbf{Q}$. Let the $\mathbf{z}$'s in (5) be a GMRF defined on a regular lattice, then in case of $\alpha = 2$ the appropriate precision matrix is obtained by

$$\mathbf{Q} = \frac{1}{\tau^2}\big(\kappa^4\mathbf{C} + 2\kappa^2\mathbf{G} + \underbrace{\mathbf{G}\mathbf{C}^{-1}\mathbf{G}}_{\mathbf{G2}}\big) \tag{9}$$

where $\mathbf{C}$, $\mathbf{G}$, and $\mathbf{G_2}$ are sparse matrices (see Lindgren et al., 2011, for details). A special case of (8) is the intrinsic Matérn model with $\kappa = 0$, given by

$$(-\Delta)^{\frac{\alpha}{2}} z(\mathbf{s}) = \tau\mathcal{W}(\mathbf{s}), \tag{10}$$

this is a spline smoothing model (see Duchon, 1976, Kimeldorf and Wahba, 1970, Nychka, 2000, Wahba, 1981). In this case, $\mathbf{Q}$ is the precision matrix of an intrinsic GMRF.

To obtain a suitable dependence model for $\mathbf{z}$, we assume the same precision $\mathbf{Q}$ for both fields, $\mathbf{z}_1$ and $\mathbf{z}_2$, in (5), but allow the fields to be correlated. The result is a separable precision (inverse covariance) matrix that can be expressed as a Kronecker product,

$$\mathbf{\Sigma_z}^{-1} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \otimes \mathbf{Q}, \tag{11}$$

where ρ is the correlation between the fields.

2.2.3 Interpretation of the statistical model

The purpose of the residual field is to capture spatial structure not explained by the mean field; if GB-REVEALS in several nearby grid cells deviate in a similar fashion from the regression model then predictions at adjacent grid cells should take this information into account. The intrinsic GMRF model (IGMRF) used here can be interpreted as either universal Kriging using a Matérn-covariance with very large range (Lindgren et al., 2011), or as spline smoothing of the residuals (Nychka, 2000, Wahba, 1981). Predictions at an unobserved grid location are essentially given by

$$\begin{aligned} \hat{u}_i(\mathbf{s}_0) =& \hat{\beta}_{0,i} + \sum_p B_p(\mathbf{s}_0)\hat{\beta}_{p,i} \\ &+ \sum_{j=1}^{2} \sum_{\mathbf{s}\in\text{observed}} c(\mathbf{s}-\mathbf{s}_0, j) \left(u_j(\mathbf{s}) - \hat{\beta}_{0,j} - \sum_p B_p(\mathbf{s})\hat{\beta}_{p,j} \right). \end{aligned} \tag{12}$$

Here $c(\mathbf{s}, i)$ are weighting coefficients that depend on τ, σ_ε, and ρ in (15) in A and decay as the distance from $\mathbf{s}_0$ increases (see Lindgren et al., 2011, Rue and Held, 2004, for details). Note that the predictions include residuals from both fields; this is due to the correlation, ρ, introduced in (11).

For prediction and cross-validation both the regression parameters (β) and the parameters describing the spatial structure (τ, σ_ε^2, ρ) are calibrated based on a validation set (i.e. selected grid cells with GB-REVEALS). These parameters are then used to compute predictions at unobserved sites and sites left-out for the cross-validation.

For details regarding the calibration and reconstruction see A.

2.2.4 Models used for data analysis

For reconstruction of land-cover types using the GB-REVEALS and LPJ-GUESS$_{KK10}$ data we consider a total of four models.

The two regression models, RM and RM$_{geo}$, without spatial dependencies and the two models, IGMRF and IGMRF$_{geo}$, with spatial dependencies. The two spatially dependent models are created by adding spatial residual fields, $\mathbf{z}$, to the same mean fields, (6), as in the regression models. All four models have been implemented in R (R Core Team, 2014) using the R-INLA package (Lindgren and Rue, 2013, Rue et al., 2009).

3 Land-cover type and auxiliary data

The target region of the LANDCLIM project is Europe (Fig. 2; Gaillard et al., 2010). The data used for validation and application of the statistical models proposed in the previous section consist of 1) GB-REVEALS extracted from the LANDCLIM-REVEALS database, 2) DVM LPJ-GUESS estimates of potential natural vegetation and ALCC KK10 scenario estimates, and 3) the present-day forest map of Europe, geographical coordinates (longitude and latitude), and elevation.

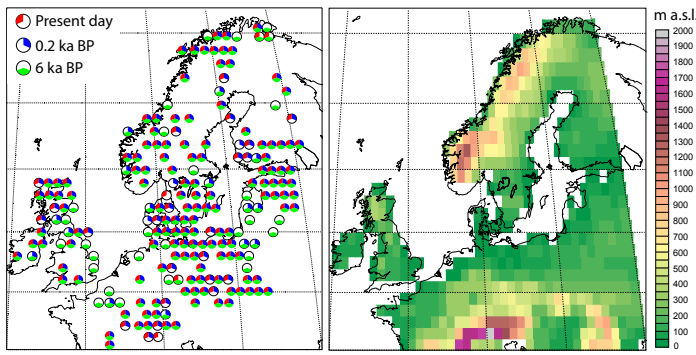


Figure 2: The left plot shows the availability of GB-REVEALS for all three time-windows. The right plot shows the elevation for each grid cell (truncated to ≥ 0).

The two selected time windows of the past are characterized by contrasting human impact on land cover (Gaillard et al., 2010, Strandberg et al., 2014) and recent land cover is used to validate the methods:

6 ka (5.7 – 6.2 ka) — the mid-Holocene warm period characterized by low human impact, often used as a baseline time-window to assess the effects of orbital forcing and pre-industrial greenhouse gases on climate (e.g. Braconnot et al., 2012, Harrison et al., 1998, Kohfeld and Harrison, 2000).

0.2 ka (0.1 – 0.35 ka) — the Little Ice Age (AD 1550 – 1850), a cool period in Europe with substantial anthropogenic land cover but low levels of human-induced greenhouse gases; AD 1850 and AD 1750 were used as pre-industrial baselines in the two last IPCC (Intergovernmental Panel on Climate Change) reports (Pachauri and Reisinger, 2007, Stocker et al., 2013).

0.05 ka (x[1]-0.1 ka) — recent land cover is characterized by afforestation of large areas of Europe in mountainous areas and other regions, such as southern Sweden, northern Germany and Poland, and the Baltic states (Krzywinski et al., 2009).

3.1 Land-cover reconstruction using fossil pollen

The LCT data used for analysis are calculated from mean REVEALS estimates for 25 major plant taxa in $1^\circ \times 1^\circ$ grid cells following the LANDCLIM protocols described in Gaillard et al. (2010), Mazier et al. (2012), Trondman et al. (2014). REVEALS (Sugita, 2007) is a mechanistic model for regional vegetation reconstruction that takes into account the inter-taxonomic differences in pollen productivity and dispersal, and size of sedimentary basins from which pollen data are obtained. It has been tested in several regions of Europe and North America (Hellman et al., 2008a,b, Soepboer et al., 2010, Sugita et al., 2010). Hellman et al. (2008b) showed that the spatial scale of REVEALS estimates of land cover was generally closer to 100 km × 100 km than 50 km × 50 km; however, the difference in fit between the REVEALS estimates and the actual land cover was small between the two spatial scales. Therefore, the $1^\circ \times 1^\circ$ spatial scale chosen for the LANDCLIM project is adequate for the REVEALS model. The LANDCLIM database includes more than 600 Holocene pollen records that are compiled from the European Pollen Database, various national pollen databases and archives, and individual contributors. In our study, the number of $1^\circ \times 1^\circ$ grid cells with GB-REVEALS are 184, 179, and 168 (out of a total of 644, 675, and 658 grid cells

[1] x = date of the core surface, e.g. AD 2005-100 BP if x = AD 2005

over the study region) for the 6 ka, 0.2 ka, and 0.05 ka time windows, respectively (Fig. 2). The total number of grid cells and the number of grid cells with GB-REVEALS differ among the time periods because of the differences in the coastline of the ALCC scenario used, and the differences in availability of pollen data among the time windows, respectively.

3.2 Estimates of potential vegetation and anthropogenic land cover

For estimating changes in the distribution and cover of LCTs in each $1^\circ \times 1^\circ$ grid cell, we use a combination of simulated potential vegetation and estimates from ALCC scenario as follows.

3.2.1 Potential natural vegetation

Potential vegetation in the study region is simulated by a process-based dynamic ecosystem model LPJ-GUESS 2.1 (Lund-Potsdam-Jena-General Ecosystem Simulator, Sitch et al., 2003, Smith et al., 2001). For the three, above specified, time windows LPJ-GUESS was run using climate input data provided by 1) the SMHI Rossby Centre Regional Climate Model (RCA3) (Samuelsson et al., 2011) with a 0.44° spatial resolution over Europe for 6 ka (5859–5811) BP and 0.2 ka (AD 1700–1800), (Strandberg et al., 2014) and 2) the Climatic Research Unit with a 0.5° resolution for 0.05 ka (AD 1901–2006) (Mitchell and Jones, 2005). The modern soil-texture data as described in Sitch et al. (2003) were used in all simulations. Percentage covers of plant functional types simulated by LPJ-GUESS are averaged over the modelled periods and converted to the three LCTs in proportion. All the estimates are upscaled, by averaging, to $1^\circ \times 1^\circ$ resolution. See B for a more detailed description of mechanisms and parameterizations in LPJ-GUESS.

3.2.2 Anthropogenic deforestation

Anthropogenic deforestation in the study region is extracted from the standard scenario of the ALCC KK10 (Kaplan et al., 2009). The KK10 scenario is based upon estimates of past human population density and the land requirement per capita to estimate the area of land needed for sustaining the assumed population. The spatial distribution of anthropogenic land cover is determined by environmental suitability estimates based mainly on climate conditions and soil type. We chose the KK10 scenario to represent anthropogenic deforestation, because there is a good correlation between GB-REVEALS and KK10 (Kaplan et al., 2014,

Strandberg et al., 2014, Trondman et al., 2012). The KK10 scenario provides estimates of the fraction of land used for agrarian activities (i.e. deforested land) at $5'$ spatial resolution. These estimates are averaged for 100-year time windows around 6 ka and 0.2 ka BP and upscaled to $1^\circ \times 1^\circ$ resolution.

3.2.3 LCT estimates based on the LPJ-GUESS and KK10 simulations

Because LPJ-GUESS does not account for the increase in unforested area due to human impact (Fig. 3), this study uses the following adjustment to estimate land cover:

$$\begin{aligned}
\mathbb{P}(\text{Coniferous}_{\text{adj.}}) &= \mathbb{P}(\text{Coniferous}) \cdot \big(1 - \mathbb{P}(\text{HLU}_{\text{KK10}})\big), \\
\mathbb{P}(\text{Broadleaved}_{\text{adj.}}) &= \mathbb{P}(\text{Broadleaved}) \cdot \big(1 - \mathbb{P}(\text{HLU}_{\text{KK10}})\big), \\
\mathbb{P}(\text{Unforested}_{\text{adj.}}) &= \mathbb{P}(\text{Unforested}) \cdot \big(1 - \mathbb{P}(\text{HLU}_{\text{KK10}})\big) + \mathbb{P}(\text{HLU}_{\text{KK10}}).
\end{aligned} \tag{13}$$

In each grid cell, the LPJ-GUESS-based proportions of the area occupied by coniferous forest, broadleaved forest and unforested land are expressed as $\mathbb{P}$(Coniferous), $\mathbb{P}$(Broadleaved) and $\mathbb{P}$(Unforested), respectively; the estimated proportion of human induced deforestation from the KK10 scenario is $\mathbb{P}(\text{HLU}_{\text{KK10}})$. Adjusted land cover proportions — $\mathbb{P}(\text{Coniferous}_{\text{adj.}})$, $\mathbb{P}(\text{Broadleaved}_{\text{adj.}})$ and $\mathbb{P}(\text{Unforested}_{\text{adj.}})$ — are calculated in such a way that the cover fractions of all the potential vegetation components are uniformly reduced by the predicted anthropogenic land-cover proportion. The anthropogenic land-cover proportions, $\mathbb{P}(\text{HLU}_{\text{KK10}})$, are then added to the unforested fraction, $\mathbb{P}$(Unforested). The resulting adjusted land-cover proportions are then used as explanatory variables in the mean field, (6).

3.3 Present-day land cover, elevation, longitude and latitude

Data on the present-day land cover in the study region were obtained from the forest map of Europe compiled by the European Forest Institute (EFI). Raster maps based on a combination of satellite data (NOAA-AVHRR) and national forest-inventory statistics from 1990–2005 (Päivinen et al., 2001, Schuck et al., 2002) were downloaded from the EFI webpage (`http://www.efi.int/portal/virtual_library/information_services/mapping_services/forest_map_of_europe`). The forest maps (with proportions of coniferous- and broadleaved-forest cover) were upscaled, by averaging, from 1 km $\times$ 1 km to

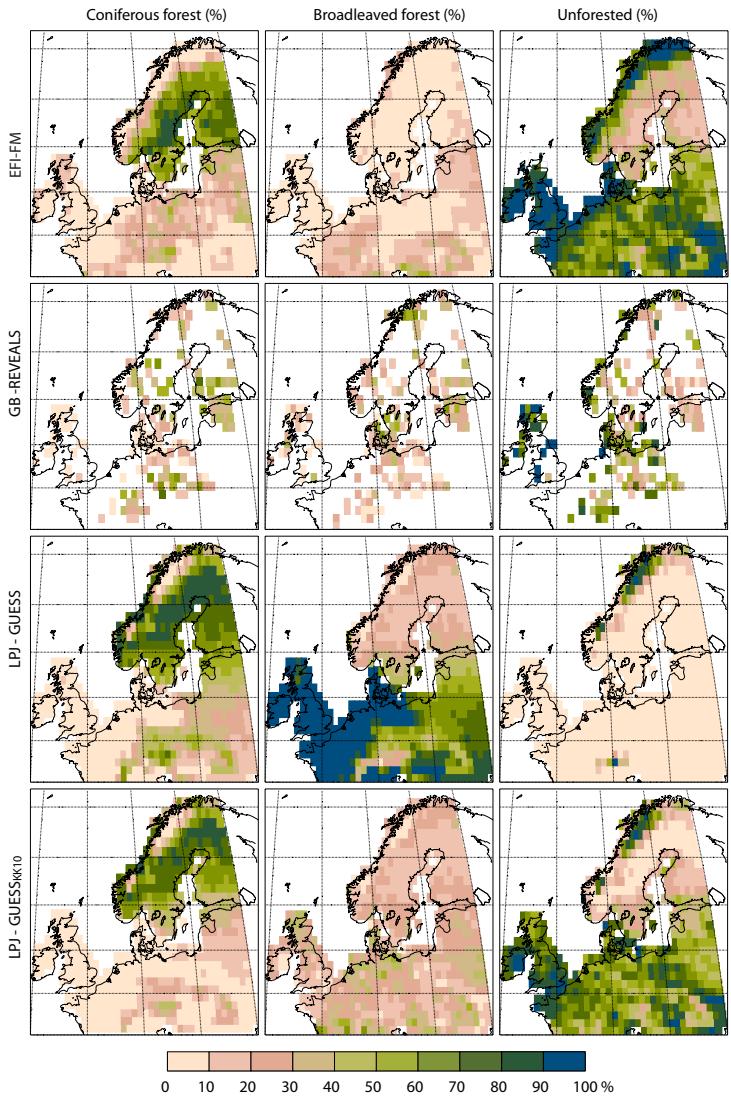


Figure 3: Available data for the 0.05 ka period, showing the proportion of LCTs. From top to bottom, data for the present-day window from EFI-FM, GB-REVEALS, LPJ-GUESS, and LPJ-GUESS$_{KK10}$ ((13)).

$1^\circ \times 1^\circ$ resolution. The proportion of unforested area were calculated by subtracting the total sum of forested covers from 1.0.

The elevation data were obtained from the Shuttle Radar Topography Mission (SRTM) (Becker et al., 2009) downloaded from `ftp://topex.ucsd.edu/pub/srtm30_plus/` on 2011–09–03, averaged over each GB-REVEALS grid-cell, and truncated to ≥ 0 to avoid a few grid cells along Norway's coast with elevation down to -1000. The geographical coordinates consist of the longitude and latitude of the central point of each GB-REVEALS grid cell.

4 Results

4.1 Evaluation of the statistical models

To evaluate and validate the four statistical models, we compared the differences between the reconstructed values for the 0.05 ka and the data from the EFI forest map (EFI-FM) using compositional distances ((4)) for individual grid cells. The average distances (i.e. the mean difference between the model-based reconstruction and EFI-FM) are 1.711, 1.520, 1.782, and 1.517 for $\text{IGMRF}_{\text{geo}}$, IGMRF, RM_{geo}, and RM, respectively. Thus the models without geographic coordinates (RM and IGMRF) provide estimates closer to those from EFI-FM than the other two models.

We also adopted a 6-fold cross-validation scheme for each of the three time windows (see Hastie et al., 2001, ch. 7.10). To assess the possible variability due to the selection of different groupings, the cross-validation is run for 10 different, randomly selected, 6 folds. Average compositional errors and standard deviations are shown in Table 1. For the cross-validation (CV) the RM model is consistently best for all time windows followed by RM_{geo}, IGMRF, and $\text{IGMRF}_{\text{geo}}$. However, longitudinal effects introduced by the geographic coordinates result in unsatisfactory reconstructions from the RM_{geo} and $\text{IGMRF}_{\text{geo}}$ model for areas of eastern Europe with few GB-REVEALS. Due to the scarcity of the GB-REVEALS data, these longitudinal effects are not penalised by the cross-validation, but show up in the comparisons with EFI-FM data; this leads us to prefer IGMRF over the RM_{geo}.

In addition, we computed the deviance information criteria (DIC; see Ch. 7.2 Gelman et al., 2014) for each of the models (Table 2); the DIC is a generalization of the Akaike information criterion (AIC; Akaike, 1969). The DIC suggests that the IGMRF models outperform the regression models for all time windows,

indicating the need for spatial dependency. The results of CV and DIC lead us to choose RM and IGMRF as the best model.

Accordingly we proceeded with the data analysis using the RM and IGMRF only. The reconstructions from the RM and IGMRF are shown in Fig. 4.

	0.05 ka		0.2 ka		6 ka	
Model	CV_{error}	(sd)	CV_{error}	(sd)	CV_{error}	(sd)
RM	1.565	(0.033)	1.843	(0.057)	1.761	(0.041)
RM_{geo}	1.631	(0.051)	1.985	(0.068)	1.825	(0.049)
IGMRF	1.679	(0.046)	2.020	(0.071)	1.928	(0.048)
$IGRMF_{geo}$	1.705	(0.049)	2.060	(0.077)	1.956	(0.053)

Table 1: Average compositional error (and standard deviation) from 10 different 6-fold cross-validations for each of the 4 different models, and 3 time windows.

DIC	0.05 ka	0.2 ka	6 ka
$IGMRF_{geo}$	750.13	891.84	-2123.6
IGMRF	664.05	839.51	-2128.82
RM_{geo}	1058.59	1194.41	1319.18
RM	1142.96	1348.56	1372.9

Table 2: Deviance information criteria (DIC) for each of the 4 different models, and 3 time windows.

4.2 Assessment of the data quality

To gain an overall understanding of data quality and to detect possible method-inherent biases, the compositional distances are calculated between the EFI-FM and either LPJ-$GUESS_{KK10}$, GB-REVEALS, or the statistical reconstructions, RM and IGMRF, for the 0.05 ka. The distances between LPJ-$GUESS_{KK10}$ and GB-REVEALS are also computed for each of the three time windows, allowing us to investigate how much these two datasets differ in each time window. The discrepancy between LPJ-$GUESS_{KK10}$ and GB-REVEALS increases from the 0.05 ka to 6 ka (average distance of 1.644, 1.701, and 2.054 for the 0.05 ka, 0.2 ka, and 6 ka windows, respectively; Fig. 5). Reasons behind these discrepancies are presented in Sec. 5.2.

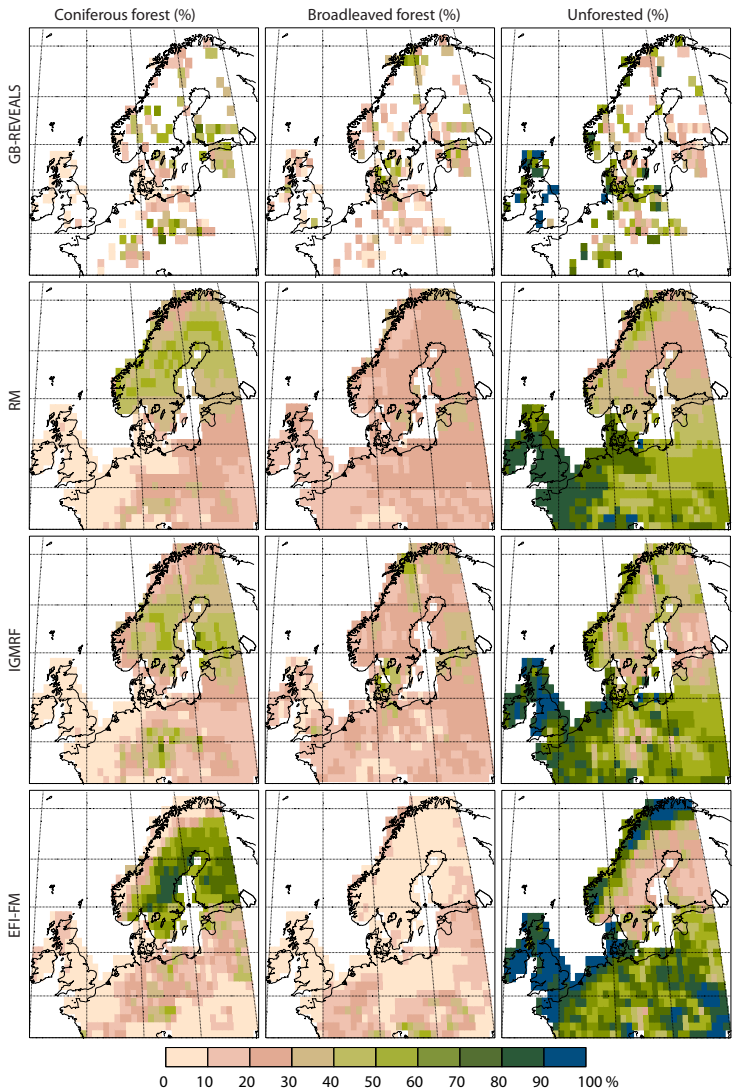


Figure 4: Reconstructions of proportion of LCTs for the 0.05 ka time window. From top to bottom, the REVEALS data, the RM reconstruction, the IGMRF reconstruction and the EFI-FM data.

For the 0.05 ka time window the RM- and IGMRF-based reconstructions are closer to EFI-FM (average distance: 1.517 and 1.519, respectively) than either LPJ-GUESS$_{KK10}$ (2.081) or GB-REVEALS (1.675) alone. Further, the RM- and IGMRF-based reconstructions, when averaged *only over grid cells* where we have GB-REVEALS, are closer to the EFI-FM data than GB-REVEALS — RM 1.499, IGMRF 1.592, and GB-REVEALS 1.675. Thus, the statistical modelling approach reduces the compositional error of the land-cover reconstructions.

Spatially, discrepancies between EFI-FM and GB-REVEALS or LPJ-GUESS$_{KK10}$ are noticeable along the coastal areas of western Europe, in central Sweden and southern Finland, with the largest disagreement at the northern British Isles. The discrepancies are more pronounced for LPJ-GUESS$_{KK10}$ than either GB-REVEALS or one of the proposed statistical model (Fig. 5). When comparing LPJ-GUESS$_{KK10}$ and GB-REVEALS, the patchy nature of GB-REVEALS makes it hard to distinguish any specific spatial patterns among the discrepancies.

4.3 Qualitative differences among the models

In general, IGMRF captures more of the local variability in GB-REVEALS than RM, while RM smooths GB-REVEALS. For example, the high variation of unforested land in Britain and the Alps is well captured by IGMRF while RM is capable of capturing the gradual changes in vegetation abundances observable along the western coast of Norway and around the northern Baltic (Fig. 4).

Both statistical models overestimate the abundance of broadleaved forest relative to the EFI-FM data. Figure 6 shows that the low abundance of broadleaved forest in EFI-FM can be seen as a cluster of EFI-FM data along the U-C edge (corresponding to $\sim$ 0% of broadleaved cover) for which no matching GB-REVEALS exists. This is due to GB-REVEALS having a higher abundance of broadleaved forest than EFI-FM (Fig. 6).

4.4 Statistical reconstruction of past land cover

Figures 7 and 8 show the reconstructed land cover at 0.2 ka and 6 ka using RM and IGMRF. In northern Europe, there is a general shift from largely unforested to more coniferous-dominated land cover between 0.2 ka and 0.05 ka (Fig. 4). The shift reflects a considerable decrease in agrarian land use in favour of modern forestry with conifer species in many regions (e.g. Fredh et al., 2013, Poska et al., 2008). Both models capture the shift (Figs. 4 and 7). Further, IGMRF is

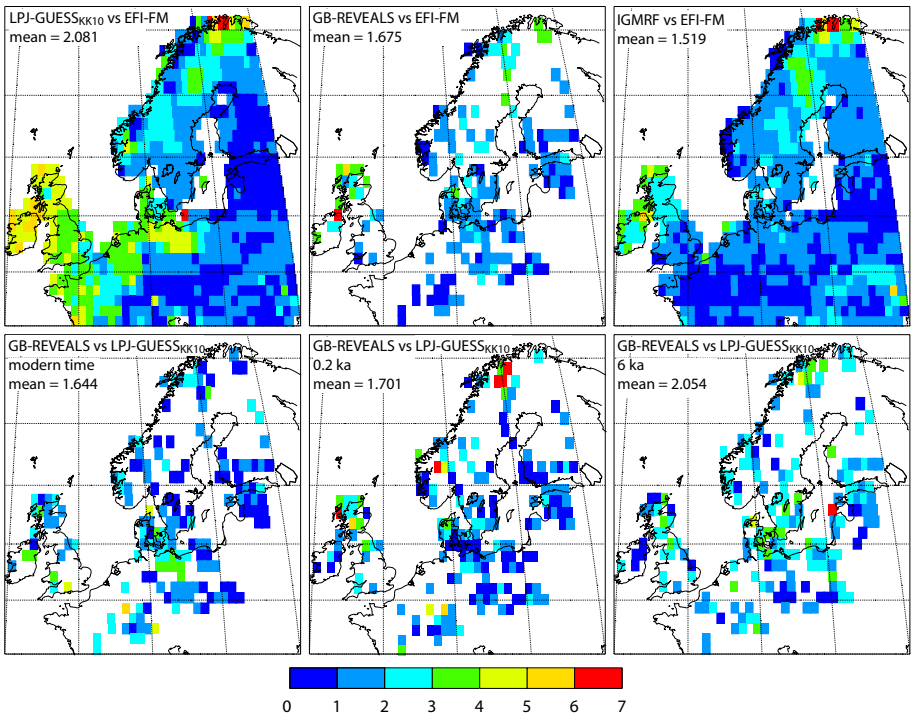


Figure 5: The compositional distances among different datasets. The first row shows the distances between EFI-FM, and (from left to right) 1) LPJ-GUESS$_{KK10}$, 2) GB-REVEALS, and 3) one of the proposed models, IGMRF for the 0.05 ka time window. The second row shows average distances between GB-REVEALS and LPJ-GUESS$_{KK10}$ for three time-windows, 0.05 ka, 0.2 ka, and 6 ka.

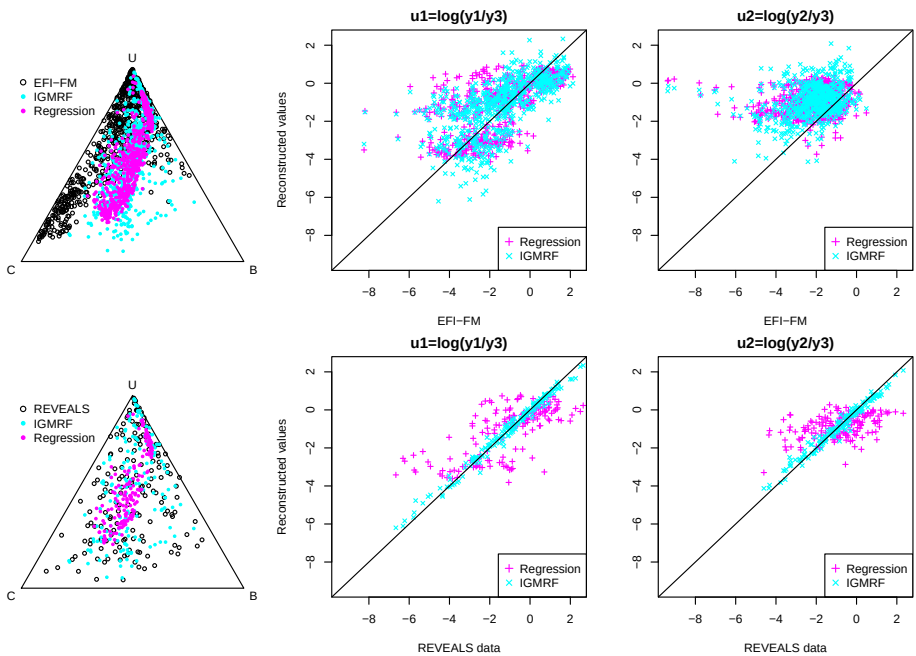


Figure 6: The reconstructions from RM and IGMRF against the EFI-FM data on the top row, and against the 0.05 ka GB-REVEALS on the bottom. The first column shows the ternary diagram for three compositions (C: Coniferous forest, B: Broadleaved forest, and U: Unforested land). Note the cluster of data points along the U-C edge, representing the low abundance of broadleaved forest in the EFI-FM data which does not exist in the 0.05 ka GB-REVEALS. The remaining columns show the scatter plots of alr-transformed, (u_1, u_2), reconstructed values for IGMRF and RM.

capable of capturing a number of sub-regional structures such as the high abundance of unforested land in Britain, especially in its northernmost regions (Fyfe et al., 2013) and the high abundance of coniferous forest around the Alps, central Sweden, and southern Finland; these patterns are also present in GB-REVEALS for 0.2 ka.

Broadleaved forest is a major constituent of the land cover at 6 ka over Europe (Fig. 8). IGMRF captures the GB-REVEALS structure with a locally highly varying (between 20% and 80%) abundance of unforested land in western Europe and the Carpathians, while the RM produces a smoother reconstruction with an averaged (around 40% to 50%) and regionally smooth abundance of unforested land. IGMRF also captures the higher than average abundance of coniferous forest in the south-eastern Baltic states and the Alps, while RM only captures these features around the Alps.

Statistical reconstructions of land cover for the 0.05 ka and 0.2 ka time windows show a good fit of the statistical models to the EFI-FM data (for 0.05 ka) and to GB-REVEALS (for both times). For the 6 ka time window the increasing discrepancies between GB-REVEALS and LPJ-GUESS$_{KK10}$ makes it hard for the statistical models to combine the two data sets; resulting in either over-smoothing (RM) or exaggeration (IGMRF) of local structures in the GB-REVEALS.

5 Discussion

5.1 Advantage of the new approach over previously proposed methods

The statistical method developed and used in this paper utilizes GMRFs (Lindgren et al., 2011) and the R-INLA package (Lindgren and Rue, 2013, Rue et al., 2009) to obtain fast inference for a complex statistical model using standard tools. Spatially dependent compositional data has previously been modelled using similar approaches with different specifications of the latent field, different applications, and more time consuming calibration methods (e.g. Billheimer et al., 2001, Paciorek and McLachlan, 2009, Tjelmeland and Lund, 2003).

In addition to the statistical methods used, the inclusion of estimates from LPJ-GUESS and KK10 provides a way of capturing the non-linear effects of bioclimatic variables on land cover by combining data from a DVM (LPJ-GUESS) and an ALCC scenario (KK10) with grid-based pollen estimates (GB-REVEALS). All the suggested models rely, to some extent, on covariates and to obtain a good

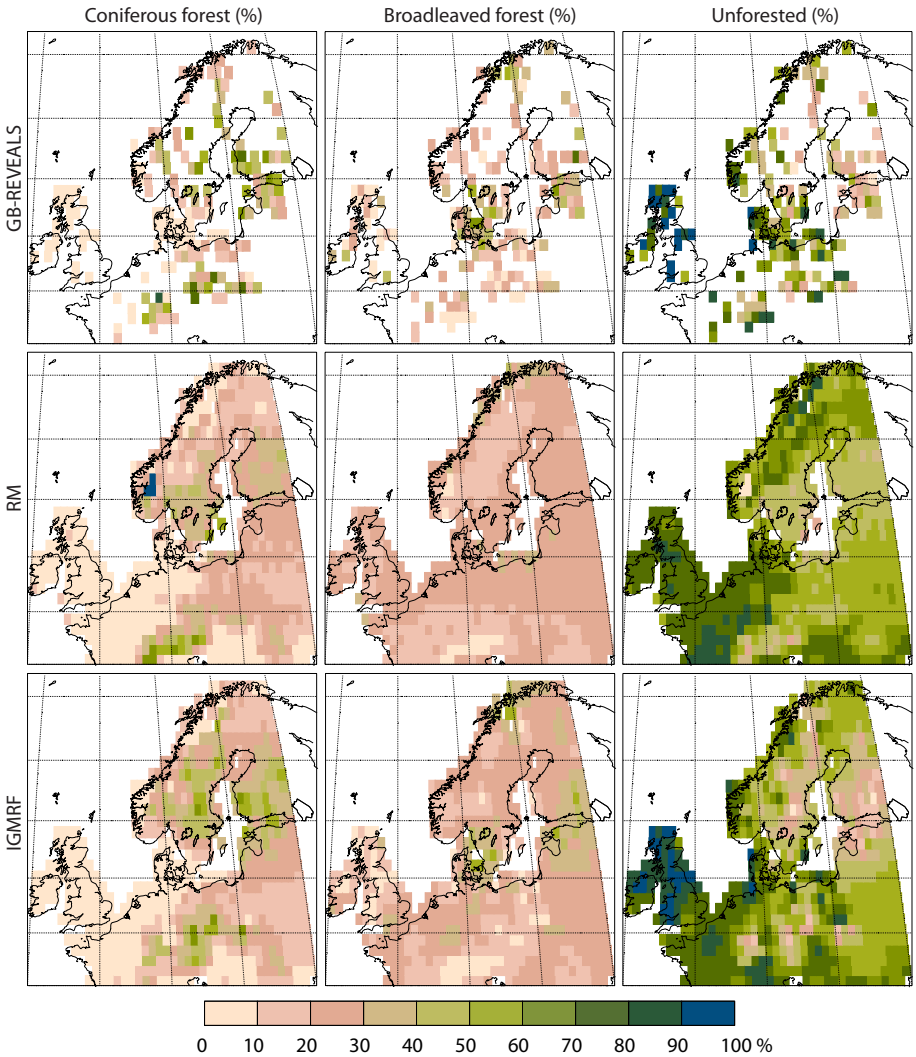


Figure 7: Reconstructions for the 0.2 ka time window of proportion of LCTs. From top to bottom, GB-REVEALS, the RM reconstruction, and the IGMRF reconstruction.

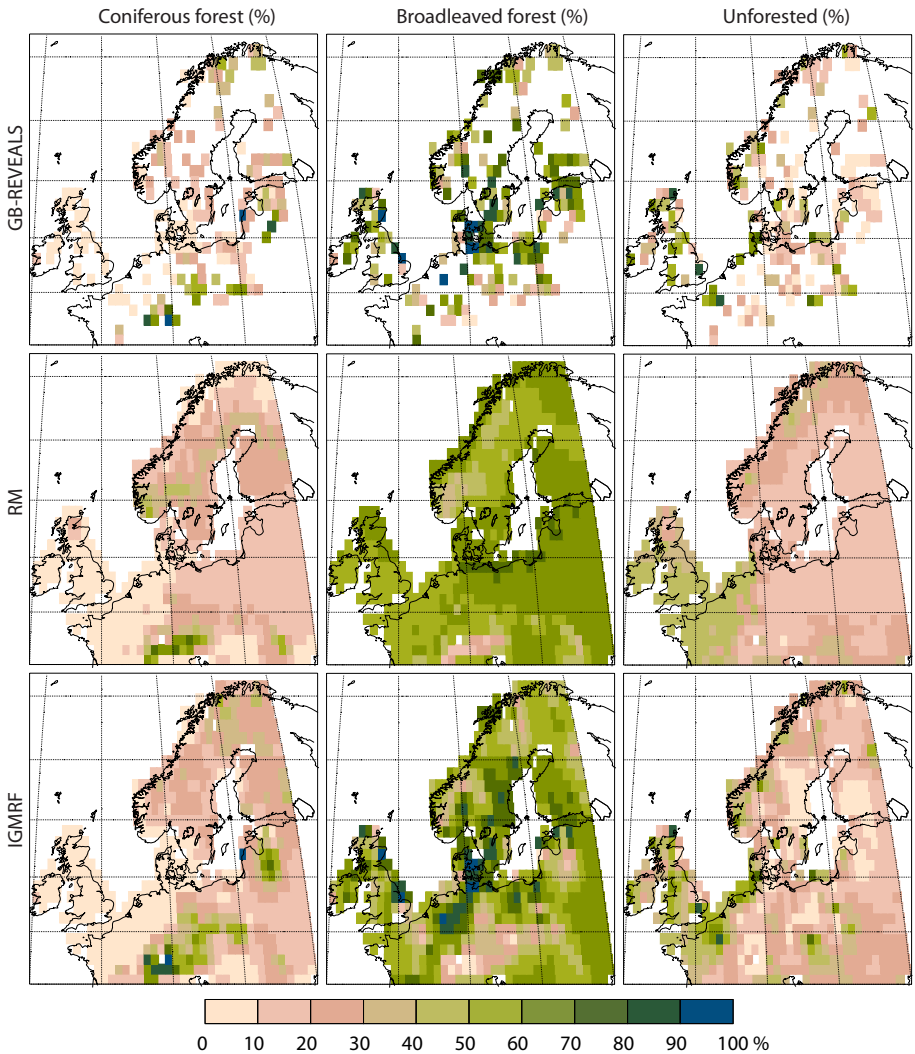


Figure 8: Reconstructions for the 6 ka time window of proportion of LCTs. From top to bottom, GB-REVEALS, the RM reconstruction, and the IGMRF reconstruction.

reconstruction it is important to identify and use covariates with strong explanatory power. For regions with few GB-REVEALS the statistical modelling will essentially extrapolate the covariate behaviour observed in other regions with more data. For our study the DVM based covariates produce good reconstructions, while the geographic covariates in the RM_{geo} model lead to a reconstruction exhibiting longitudinal effects that do not correspond to patterns in the EFI-FM data or in the GB-REVEALS.

Pollen-inferred land cover has also been studied by Paciorek and McLachlan (2009), but over a much smaller geographic area (roughly equivalent to one of the grid cells in our approach) than our continent-wide focus. Paciorek and McLachlan (2009) combined pollen data with maps of vegetation abundances to produce a combined estimate of pollen productivity and land cover. Thus the focus and geographic extent of their paper is closer to that of REVEALS (Sugita, 2007) than to ours. Another potential use of the pollen data (Garreta et al., 2010) is to attempt recovery of past climate by the inversion of a dynamic vegetation model; this approach provides past climate, but does not provide spatial reconstructs of the past land-use.

Although the model provides good results, it may be improved, for instance by introducing a more elaborate method of adjusting potential vegetation for human deforestation, (13) than that used in this study. Currently, we assume that human deforestation affects the three land cover classes equally. An interesting extension would be to include, and estimate, the differential impact of deforestation on the natural vegetation types. Another potential improvement would be the inclusion of the estimated uncertainties in GB-REVEALS; this might allow the model to disregard uncertain GB-REVEALS values, possibly improving the reconstruction.

5.2 Biases of land-cover databases and their effects on RM and IGMRF applications

5.2.1 GB-REVEALS

Possible sources of errors and biases in the GB-REVEALS are discussed in details in Trondman et al. (2014). Here we mention the most important ones. The pollen productivity estimates (PPEs) used to obtain the GB-REVEALS are based on pollen and vegetation data from low-land areas of NW and W Europe (Broström et al., 2008, Mazier et al., 2012). This might lead to biased estimates of the regional vegetation in areas where region-specific PPEs are not available,

such as the high mountain areas of Norway, the Czech Republic and the Alps. Also, the use of all pollen records available in each grid cell, i.e. pollen data from both lakes and bogs (small or large in size) in the LANDCLIM-REVEALS database, may create errors caused by variations in the number, type and size of the sites used to calculate the GB-REVEALS (Mazier et al., 2012, Trondman et al., 2014). The best pollen records for the application of REVEALS are those from large lakes (one or several, $>$ 20–30 ha) or multiple small lakes ($\leq$ 20–30 ha; see Hellman et al., 2008a, Sugita, 2007, for details). If REVEALS is applied on pollen data from small sites only, a small number of sites may result in very large error estimates (Mazier et al., 2012, Sugita, 2007, Trondman et al., 2014). Further, the REVEALS model assumes that no vegetation is growing on the surface of the basin where pollen is deposited (Sugita, 2007), which applies to lakes only. Although REVEALS includes two versions of the pollen dispersal and deposition model (one for lakes and one for bogs) the assumption mentioned above is violated. Therefore, pollen data from large bogs in particular might bias the GB-REVEALS due to the local vegetation on the bog (Trondman et al., 2014).

In addition, locally grown shade-intolerant deciduous trees, such as *Alnus* and *Betula*, on the wetland and along the shores of the lakes tend to be over-represented in pollen records. This may bias the REVEALS reconstructions because the model cannot fully correct for the over-representation. Although broadleaved trees are over-represented compared to EFI-FM in southern Sweden in this study, the validation of the REVEALS model in the same areas do not exhibit any such over-representation (Hellman et al., 2008a,b).

When comparing EFI-FM with the 0.05 ka GB-REVEALS it is important to note that: i) due to the spectral reflectance of broadleaved forest in combination with that of water resembling quite closely the spectral reflectance of coniferous forest, EFI tends to underestimate broadleaved forest along water courses (Schuck et al., 2002); ii) EFI-FM had the lowest accuracy for broadleaved forest (c.f. Section 5.2.4); and iii) the GB-REVEALS cover a much longer time interval (from AD 1850 to the year of coring at each site) than the EFI-FM (inventory and satellite data from 1990–2005) and, during the past century, parts of our study region were characterized by the abandonment of traditional agriculture in favour of sylviculture with plantations of *Picea* and *Pinus* (e.g. Cui et al., 2014, Fredh et al., 2013, Krzywinski et al., 2009, Poska et al., 2008).

Any biases in GB-REVEALS will be propagated into the statistical reconstructions and it is not possible to assess the detailed effects of such biases on the RM

and IGMRF applications neither in qualitative nor in quantitative terms. But it is important to have these possible biases in mind when the RM and IGMRF reconstructions are discussed and validated against present-day data. Nevertheless, GB-REVEALS used in this study are mostly credible for the time interval they represent (Trondman et al., 2014) and any biases should be small.

5.2.2 LPJ-GUESS

The largest discrepancies between LPJ-GUESS simulated land cover and present-day EFI-FM or palaeo-based GB-REVEALS of past land cover coincide with areas characterized by high rainfall and/or long-term anthropogenic land cover and inherent specific land-cover types such as heathlands or blanket-bogs that are difficult to model using a natural terrestrial vegetation models. For instance, the LPJ-GUESS standard soil biogeochemistry used in this study, which excludes the nutrient cycle (Sitch et al., 2003), can lead to imprecise estimates of vegetation composition in nutrient-limited environments, for example at high latitudes (Wårlind, 2013). Moreover, the absence of dispersal and migratory processes in the LPJ-GUESS standard setup (Smith et al., 2001) leads to an overrepresentation of coniferous forest (spruce in particular) in central and northern Europe, especially during the early and mid Holocene (Lehsten et al., 2014), which may affect the land-cover reconstruction at 6 ka. Post-processing of the LPJ-GUESS simulated natural vegetation for migration processes of taxa such as *Fagus* (beech) may decrease the difference between LPJ-GUESS and GB-REVEALS at 6k (Poska et al., 2012).

At 0.05 ka (Fig. 3) there is a similar discrepancy between LPJ-GUESS$_{KK10}$ and GB-REVEALS along the coasts of Norway. Here, LPJ-GUESS simulates high cover of coniferous forest while GB-REVEALS exhibit higher cover of broadleaved forest. This is a consequence of the bias in LPJ-GUESS mentioned above, but also of long-term human impact in these regions with the development of grazed heaths from the Neolithic time that still cover large areas (e.g. Gaillard et al., 2009).

Further, the LPJ-GUESS estimates of land-cover composition are highly dependent on the climate input data (RCA3 simulations of past climate) used to force LPJ-GUESS. Strandberg et al. (2014) showed that there are some discrepancies between proxy-based reconstructions of past climate and RCA3 simulations; although both climate palaeo-proxies and RCA3 simulations show higher temperatures at 6 ka than at 0.2 ka, the difference in magnitude between the two time

windows and the geographical/spatial patterns of reconstructed versus simulated temperature and precipitation can be very large. Biases are seen in particular in the Scandinavian mountains and in eastern and north-eastern Europe. The latter will in turn bias the LPJ-GUESS simulated vegetation and, therefore, the RM and IGMRF applications.

5.2.3 ALCC KK10

The ALCC scenario used for adjustment of LPJ-GUESS is based on the following assumptions: 1) the parameters that drive deforestation are similar in different population regions, 2) the areas with highest suitability for farming are deforested first, and 3) agricultural products were the major food source for human populations (Kaplan et al., 2009). As the extent and intensity of population pressure on the landscape may be characterized by strong regional to local-scale spatial and temporal differences in terms of technology development and usage of non-agricultural food resources, these assumptions might cause over- or underestimations of deforestation, especially for the far past (here 6 ka). For instance, the high fractions of deforested land in the ALCC scenario at 6 ka in southern Sweden and Belgium do not seem reasonable when compared to GB-REVEALS at individual sites in e.g. southern Sweden (Cui et al., 2013, Gaillard et al., 2010).

Further, the low fraction of deforested land along the coasts of Norway in the ALCC scenarios at 0.05 ka does not agree with the cover of unforested land in EFI-FM and GB-REVEALS. The ALCC scenario underestimates unforested land in these areas because the geographical and geological characteristics do not correspond to conditions associated with good suitability for farming. As a consequence, the LPJ-GUESS$_{KK10}$ estimates of deforested land may bias the RM and IGMRF applications. Moreover, the correction of the LPJ-GUESS estimates with the ALCC scenario assumes that all three LCTs are equally suitable for human land use, which is not necessarily the case. Many archaeological and palaeoecological studies in Europe have shown that the areas covered by deciduous forests tended to be deforested first for cultivation and grazing because of the favourable soil conditions (e.g. Gaillard and Göransson, 1991, Poska et al., 2004). The latter could also bias the RM and IGMRF results.

5.2.4 EFI-FM

The quality assessment of the EFI-FM by Kempeneers et al. (2012) shows 88% overall accuracy of the dataset, with accuracy for broadleaved forest being the lowest at 58%. The mapping performance was found to be spatially varying, with the best fit to ground observations in central Europe and an underestimation of tree cover in areas of sparse forest cover in Spain, Ireland and parts of Finland. This, together with the temporal miss-alignment between EFI-FM and GB-REVEALS discussed in Sec. 5.2.1, implies that model comparisons at the 0.05 ka time window needs some caution.

5.3 Implications of the results

The RM and IGMRF models show a potential to provide spatially more explicit and realistic reconstruction of the Holocene land cover than LPJ-GUESS, ALCC KK10 or REVEALS do alone.

The balance between relying on covariates (i.e. RM and mean field in IGMRF) or on nearby observations (i.e. spatial dependency part in IGMRF) is an issue in spatial statistical reconstructions. The RM model essentially consists of a regression of GB-REVEALS onto covariates (LPJ-GUESS$_{\mathrm{KK10}}$ and elevation). Large spatially varying discrepancies between GB-REVEALS and the covariates can result in an inadequate mean field, which needs to be compensated through spatial dependencies. For the 6 ka time window this is evident in the very smooth reconstructions from the RM, and the overfitting of IGMRF to GB-REVEALS. It is important to note that the RM primarily captures the large-scale variability in land cover, while IGMRF mainly captures details on a regional scale. Credible IGMRF reconstructions obviously require that the GB-REVEALS point data are reliable and that deviations from the mean model, (6), are spatially smooth. For areas with few GB-REVEALS (e.g. the northern Baltic region at 6 ka in our study, Fig. 8) the scarce data may provide a too strong local influence on the IGMRF reconstruction. In such cases, the RM reconstructions will be safer to use because individual GB-REVEALS play a less important role in the local statistical reconstruction.

6 Conclusions

The results presented here suggest that it is possible to statistically combine pollen-based reconstructions of land cover with simulated potential land cover and ALCC scenario to create spatially-explicit estimates of past land cover over large areas, such as Europe. Accurate estimates of past land cover is important, allowing for the assessment of biogeophysical effects of vegetation and land-use changes on past climate.

The proposed best models provide good reconstructions for the 0.05 ka and 0.2 ka time windows, although highlighting slightly different features. The larger differences among GB-REVEALS, LPJ-GUESS$_{KK10}$, and the statistical reconstructions at 6 ka suggest that further modifications and developments of the models are necessary to improve the estimates of land cover in older time periods. Future improvements may be possible by: 1) using a more flexible way of combining an ALCC scenario with estimates from a DVM, i.e. accounting for the varying suitability of land-cover types for agrarian activities, 2) including the error estimates of GB-REVEALS in the statistical modelling.

These pollen-based, spatially continuous land-cover reconstructions can then be used in the analysis of landscape ecological complexity in time and space (particularly the IGMRF) and in climate simulations (preferably RM) following e.g. the same scheme as Strandberg et al. (2014). A similar approach can be applied in other parts of the world, such as China, India and Africa, where long and extensive human activities have modified the earth surface significantly.

Acknowledgement

The research presented in this paper is a contribution to the two Swedish strategic research areas Biodiversity and Ecosystems in a Changing Climate (BECC), and ModElling the Regional and Global Earth system (MERGE), as well as the LAND Cover-CLIMate interactions in NW Europe during the Holocene (LANDCLIM) coordinated by M.-J. Gaillard and sponsored by the Swedish Research Council (VR), the Nordic Council of Ministers (NOrdForsk), MERGE, and the Faculty of Life and Health Sciences of Linnaeus University. We thank all LANDCLIM members who contributed pollen data to the research presented in this paper.

B. Pirzamanbein has received support from Stiftelsen Walter Gyllenbergs fund.

S. Sugita was supported by Estonian Mobilitas Programme (MTT3) and King

Carl XVI Gustaf's Foundation for Environmental Sciences in Sweden.

A Calibration and reconstruction

Parameter calibration for the models is either accomplished through standard linear regression (the RM and RM_{geo} models), or by maximising the resulting Gaussian likelihood using the R-INLA package (Lindgren and Rue, 2013, Rue et al., 2009) (the IGMRF and $\text{IGMRF}_{\text{geo}}$ models). Given calibrated parameters the transformed compositions $\mathbf{u}$ at unobserved locations are reconstructed and back-transformed (3) to obtain compositional values at all locations. Both parameter calibration and reconstruction uses the same calibration set.

For the linear regression cases the reconstruction of the alr transformed compositions at an unobserved location, $\mathbf{s}_0$, is obtain as

$$\hat{u}_i(\mathbf{s}_0) = \hat{\mu}_i(\mathbf{s}_0) = \hat{\beta}_{0,i} + \sum_p \mathbf{B}_p(\mathbf{s}_0)\hat{\beta}_{p,i}, \tag{14}$$

where $\hat{\beta}_{p,i}$ are standard linear regression estimates (i.e. parameter calibration). For the IGMRF models the reconstruction at all locations is given by

$$\begin{bmatrix}\hat{\mathbf{u}}_1\\ \hat{\mathbf{u}}_2\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{\mu}}_1\\ \hat{\boldsymbol{\mu}}_2\end{bmatrix} + \left(\boldsymbol{\Sigma}_{\mathbf{z}}{}^{-1} + \sigma_\varepsilon^{-2}\mathbf{A}^\top\mathbf{A}\right)^{-1}\sigma_\varepsilon^{-2}\mathbf{A}^\top\left(\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} - \mathbf{A}\begin{bmatrix}\hat{\boldsymbol{\mu}}_1\\ \hat{\boldsymbol{\mu}}_2\end{bmatrix}\right), \tag{15}$$

where $\hat{\boldsymbol{\mu}}_1$ is the reconstruction due to the mean field in (6) and the second term adjusts nearby locations for deviations between observations and mean-model (recall that the $\mathbf{A}$-matrix extracts the observed locations).

B LPJ-GUESS

The vegetation is simulated as plant functional types (PFTs) discriminated in terms of bioclimatic limits, growth form, phenology, life-history strategy, and various aspects of physiology. The bioclimatic niche parameterization is based on current vegetation distribution (Hickler et al., 2012). The model was run in cohort mode, in which all individuals belonging to the same age class of a PFT within a patch (local neighbourhood of individuals) are assumed to be identical in size, form, and response to the microenvironment. Multiple patches are simulated to encompass variability across the landscape of a grid cell in stand history, depending on disturbances, which recur stochastically with an expected local return time of 100 years, and stand demography. Competition for resources (light, water, etc.) among individuals is defined by the prescribed characteristics of the

PFTs in combination with the emergent vegetation structure of a patch and its effect on the microenvironment and resource availability experienced by plants. A full description of LPJ-GUESS is provided in Smith et al. (2001) and references therein. Plant physiological and ecosystem biogeochemical processes are modelled as in LPJ-DGVM (Sitch et al., 2003). The current version includes the updates described in Gerten et al. (2004) and Hickler et al. (2012).

The simulated PFT-specific leaf-area index (LAI (PTF)) output was averaged over the modelled period and converted to fractional plant cover (FPC (PTF)). The LAI (PTF) to FPC (PTF) conversion was performed by applying the Lambert-Beer law (Monsi and Saeki, 1953) to the area of ground covered by foliage directly above it (Sitch et al., 2003):

$$\mathrm{FPC(PFT)} = 1 - \exp\left(- k \cdot \mathrm{LAI(PFT)}\right)$$

References

J. Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., 1986.

H. Akaike. Fitting autoregressive models for prediction. 21:243–247, 1969.

J. J. Becker, D. T. Sandwell, W. H. F. Smith, J. Braud, B. Binder, J. Depner, D. Fabre, J. Factor, S. Ingalls, S. H. Kim, R. Ladner, K. Marks, S. Nelson, A. Pharaoh, G. Sharman, R. Trimmer, J. VonRosenburg, G. Wallace, and P. Weatherall. Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Mar. Geod.*, 32(4):355–371, 2009.

D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *J. Am. Statist. Assoc.*, 96(456):1205–1214, 2001.

P. Braconnot, S. P. Harrison, M. Kageyama, P. J. Bartlein, V. Masson-Delmotte, A. Abe-Ouchi, B. Otto-Bliesner, and Y. Zhao. Evaluation of climate models using palaeoclimatic data. *Nature Clim. Change*, 2(6):417–424, 2012.

A. Broström, A. B. Nielsen, M.-J. Gaillard, K. Hjelle, F. Mazier, H. Binney, J. Bunting, R. Fyfe, V. Meltsov, A. Poska, et al. Pollen productivity estimates of key European plant taxa for quantitative reconstruction of past vegetation: a review. *Veg. Hist. Archaeobot.*, 17(5):461–478, 2008.

V. Brovkin, J. Bendtsen, M. Claussen, A. Ganopolski, C. Kubatzki, V. Petoukhov, and A. Andreev. Carbon cycle, vegetation, and climate dynamics in the Holocene: Experiments with the CLIMBER-2 model. *Global. Biogeochem. Cy.*, 16 (4):1139, 2002.

V. Brovkin, M. Claussen, E. Driesschaert, T. Fichefet, D. Kicklighter, M. Loutre, H. Matthews, N. Ramankutty, M. Schaeffer, and A. Sokolov. Biogeophysical effects of historical land cover changes simulated by six Earth system models of intermediate complexity. *Clim. Dynam.*, 26(6):587–600, 2006.

N. Christidis, P. A. Stott, G. C. Hegerl, and R. A. Betts. The role of land use change in the recent warming of daily extreme temperatures. *Geophys. Res. Lett.*, 40(3):589–594, 2013.

M. Claussen, V. Brovkin, and A. Ganopolski. Biogeophysical versus biogeochemical feedbacks of large-scale land cover change. *Geophys. Res. Lett.*, 28(6):1011–1014, 2001.

Q. Cui, M.-J. Gaillard, G. Lemdahl, L. Stenberg, and S. Sugita. Historical land-use and landscape chaneg in southern Sweden and implications for present and future biodiversity. Submitted to Ecology and Evolution, 2014.

Q.-Y. Cui, M.-J. Gaillard, G. Lemdahl, S. Sugita, A. Greisman, G. L. Jacobson, and F. Olsson. The role of tree composition in Holocene fire history of the hemiboreal and southern boreal zones of southern Sweden, as revealed by the application of the landscape reconstruction algorithm: Implications for biodiversity and climate-change issues. *The Holocene*, 23(12):1747–1763, 2013.

N. de Noblet-Ducoudré, J.-P. Boisier, A. Pitman, G. Bonan, V. Brovkin, F. Cruz, C. Delire, V. Gayler, B. van den Hurk, P. Lawrence, M. K. van der Molen, C. Müller, C. H. Reick, B. J. Strengers, , and A. Voldoire. Determining robust impacts of land-use-induced land cover changes on surface climate over North America and Eurasia: results from the first set of LUCID experiments. *J. Climate*, 25(9):3261–3281, 2012.

J. Duchon. Splines minimizing rotation invariant seminorms in Sobolev spaces. In W. Schempp and K. Zeller, editors, *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer-Verlag, 1976.

D. Fredh, A. Broström, M. Rundgren, P. Lagerås, F. Mazier, and L. Zillén. The impact of land-use change on floristic diversity at regional scale in southern Sweden 600 BC—AD 2008. *Biogeosciences*, 10(5):3159–3173, 2013.

R. M. Fyfe, C. Twiddle, S. Sugita, M.-J. Gaillard, P. Barratt, C. J. Caseldine, J. Dodson, K. J. Edwards, M. Farrell, C. Froyd, et al. The Holocene vegetation cover of Britain and Ireland: overcoming problems of scale and discerning patterns of openness. *Quaternary. Sci. Rev.*, 73:132–148, 2013.

M.-J. Gaillard and H. Göransson. The bjäresjö area –vegetation and landscape through time. In B. E. Berglund, editor, *The cultural landscape during 6000 years in southern Sweden: the Ystad project*, volume 41 of *Ecological Bulletins*, pages 167–173. Wiley-Blackwell, 1991.

M.-J. Gaillard, T. Dutoit, K. Hjelle, T. Koff, and M. O'Connell. European cultural landscape — insights into origins and development. In K. Krzywinski, H. Küster, and M. O'Connell, editors, *Cultural Landscapes of Europe: Fields of Demeter, Haunts of Pan*, pages 35–44. Aschenbeck Media, 2009.

M.-J. Gaillard, S. Sugita, F. Mazier, A.-K. Trondman, A. Brostrom, T. Hickler, J. O. Kaplan, E. Kjellström, U. Kokfelt, P. Kuneš, , C. Lemmen, P. Miller, J. Olofsson, A. Poska, M. Rundgren, B. Smith, G. Strandberg, R. Fyfe, A. Nielsen, T. Alenius, L. Balakauskas, L. Barnekov, H. Birks, A. Bjune, L. Björkman, T. Giesecke, K. Hjelle, L. Kalnina, M. Kangur, W. van der Knaap, T. Koff, P. Lagerås, M. Latałowa, M. Leydet, J. Lechterbeck, M. Lindbladh, B. Odgaard, S. Peglar, U. Segerström, H. von Stedingk, and H. Seppä. Holocene land-cover reconstructions for studies on land cover-climate feedbacks. *Clim. Past.*, 6:483–499, 2010.

V. Garreta, P. Miller, J. Guiot, C. Hély, S. Brewer, M. Sykes, and T. Litt. A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model. *Clim. Dynam.*, 35(2–3):371–389, 2010. URL `http://dx.doi.org/10.1007/s00382-009-0629-1`.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2014.

D. Gerten, S. Schaphoff, U. Haberlandt, W. Lucht, and S. Sitch. Terrestrial vegetation and water balance-hydrological evaluation of a dynamic global vegetation model. *J. Hydrol*, 286(1):249–270, 2004.

S. Harrison, D. Jolly, F. Laarif, A. Abe-Ouchi, B. Dong, K. Herterich, C. Hewitt, S. Joussaume, J. Kutzbach, J. Mitchell, N. de Noblet, , and P. Valdes. Intercomparison of simulated global vegetation distributions in response to 6 kyr BP orbital forcing. *J. Climate*, 11(11):2721–2742, 1998.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

S. Hellman, M.-J. Gaillard, A. Broström, and S. Sugita. The REVEALS model, a new tool to estimate past regional plant abundance from pollen data in large lakes: validation in southern Sweden. *J. Quaternary. Sci.*, 23(1):21–42, 2008a.

S. E. Hellman, M.-j. Gaillard, A. Broström, and S. Sugita. Effects of the sampling design and selection of parameter values on pollen-based quantitative reconstructions of regional vegetation: a case study in southern Sweden using the REVEALS model. *Veg. Hist. Archaeobot.*, 17(5):445–459, 2008b.

T. Hickler, K. Vohland, J. Feehan, P. A. Miller, B. Smith, L. Costa, T. Giesecke, S. Fronzek, T. R. Carter, W. Cramer, I. Kühn, and M. T. Sykes. Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global. Ecol. Biogeogr.*, 21(1):50–63, 2012.

J. O. Kaplan, K. M. Krumhardt, and N. Zimmermann. The prehistoric and preindustrial deforestation of Europe. *Quaternary. Sci. Rev.*, 28(27):3016–3034, 2009.

J. O. Kaplan, K. M. Krumhardt, M.-J. Gaillard, S. Sugita, A.-K. Trondman, and F. Mazier. The deforestation history of northwest Europe: Evaluating anthropogenic land cover change scenarios with pollen-based landscape reconstructions. in preparation, 2014.

P. Kempeneers, F. Sedano, A. Pekkarinen, L. Seebach, P. Strobl, and J. San-Miguel-Ayanz. Pan-European forest maps derived from optical satellite imagery. *IEEE Earthzine*, 5, 2012.

G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2): 495–502, 1970.

K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Global. Ecol. Biogeogr.*, 20(1):73–86, 2011.

K. Kohfeld and S. Harrison. How well can we simulate past climates? evaluating the models using global palaeoenvironmental datasets. *Quaternary. Sci. Rev.*, 19(1):321–346, 2000.

K. Krzywinski, M. O'Connell, and H. Küster. *Cultural Landscapes of Europe: Fields of Demeter, Haunts of Pan*. Aschenbeck Media, 2009.

D. Lehsten, S. Dullinger, K. Hülber, G. Schurgers, R. Cheddadi, H. Laborde, V. Lehsten, L. François, M. Dury, and M. T. Sykes. Modelling the Holocene migrational dynamics of fagus sylvatica l. and picea abies (l.) h. karst. *Global. Ecol. Biogeogr.*, 2014.

F. Lindgren and H. Rue. Bayesian spatial and spatio-temporal modelling with R-INLA. Submitted to J. Stat. Softw., 2013. URL `http://www.math.ntnu.no/inla/r-inla.org/papers/jss/lindgren.pdf`.

F. Lindgren, R. Håvard, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc. Ser. B*, 73(4):423–498, 2011.

F. Mazier, M.-J. Gaillard, P. Kuneš, S. Sugita, A.-K. Trondman, and A. Broström. Testing the effect of site selection and parameter setting on REVEALS-model estimates of plant abundance using the Czech quaternary palynological database. *Rev. Palaeobot. Palyno.*, 187(1):38–49, 2012.

T. D. Mitchell and P. D. Jones. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, 25(6):693–712, 2005.

M. Monsi and T. Saeki. The light factor in plant communities and its significance for dry matter production. *Jpn. J. Bot.*, 14:22–52, 1953.

D. W. Nychka. Spatial-process estimates as smoothers. In M. G. A. Schimek, editor, *Smoothing and Regression: Approaches, Computation, and Application*, pages 393–424. Wiley, New York, USA, 2000.

R. Pachauri and A. Reisinger, editors. *Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2007. URL `http://www.ipcc.ch/publications_and_data/ar4/syr/en/contents.html`.

C. J. Paciorek and J. S. McLachlan. Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Am. Statist. Assoc.*, 104(486):608–622, 2009.

R. Päivinen, M. Lehikoinen, A. Schuck, T. Häme, S. Väätäinen, P. Kennedy, and S. Folving. *Combining earth observation data and forest statistics*. EuroForIns, 2001.

A. Pitman, N. de Noblet-Ducoudré, F. Cruz, E. Davin, G. Bonan, V. Brovkin, M. Claussen, C. Delire, L. Ganzeveld, V. Gayler, B. J. J. M. van den Hurk, P. J. Lawrence, M. K. van der Molen, C. Müller, C. H. Reick, S. I. Seneviratne, B. J. Strengers, and A. Voldoire. Uncertainties in climate responses to past land cover change: First results from the LUCID intercomparison study. *Geophys. Res. Lett.*, 36(14), 2009.

J. Pongratz, C. Reick, T. Raddatz, and M. Claussen. Effects of anthropogenic land cover change on the carbon cycle of the last millennium. *Global. Biogeochem. Cy.*, 23(4):GB4001, 2009.

J. Pongratz, C. Reick, T. Raddatz, and M. Claussen. Biogeophysical versus biogeochemical climate response to historical anthropogenic land cover change. *Geophys. Res. Lett.*, 37(8):L08702, 2010.

A. Poska, L. Saarse, and S. Veski. Reflections of pre- and early-agrarian human impact in the pollen diagrams of Estonia. *Palaeogeogr. Palaeocl.*, 209(1–4):37–50, 2004.

A. Poska, E. Sepp, S. Veski, and K. Koppel. Using quantitative pollen-based land-cover estimations and a spatial CA_Markov model to reconstruct the development of cultural landscape at Rouge, South Estonia. *Veg. Hist. Archaeobot.*, 17 (5):527–541, 2008.

A. Poska, B. Smith, D. Lehsten, L. Marquer, S. Sugita, and M.-J. Gaillard. Pollen-inferred quantitative reconstructions of past plant abundance for evaluation and development of dynamic vegetation models. In *IPC/IOPC 2012; SS07-O16 (411)*, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

W. F. Ruddiman. How did humans first alter global climate? *Sci. Am.*, March 2005:34–41, 2005.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2004.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Statist. Soc. Ser. B*, 71(2):319–392, 2009.

P. Samuelsson, C. G. Jones, U. Willén, A. Ullerstig, S. Gollvik, U. Hansson, C. Jansson, E. Kjellström, G. Nikulin, and K. Wyser. The rossby centre regional climate model rca3: model description and performance. *Tellus A*, 63 (1):4–23, 2011.

A. Schuck, J. van Brusselen, R. Päivinen, T. Häme, P. Kennedy, and S. Folving. Compilation of a calibrated European forest map derived from NOAA-AVHRR data. EFI Internal Report 13, EuroForIns, 2002.

S. Sitch, B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. Kaplan, S. Levis, W. Lucht, M. Sykes, K. Thonicke, and S. Venevsky. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Glob. Change Biol.*, 9(2):161–185, 2003.

B. Smith, I. C. Prentice, and M. T. Sykes. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Global. Ecol. Biogeogr.*, 10(6):621–637, 2001.

W. Soepboer, S. Sugita, and A. F. Lotter. Regional vegetation-cover changes on the Swiss Plateau during the past two millennia: A pollen-based reconstruction using the REVEALS model. *Quaternary. Sci. Rev.*, 29(3–4):472–483, 2010. URL `http://www.sciencedirect.com/science/article/pii/S0277379109003308`.

T. F. Stocker, D. Qin, G.-K. Plattner, M. M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, editors. *IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. URL `https://www.ipcc.ch/report/ar5/wg1/`.

G. Strandberg, J. Brandefelt, E. Kjellström, and B. Smith. High-resolution regional simulation of last glacial maximum climate in Europe. *Tellus. A*, 63(1): 107–125, 2011.

G. Strandberg, E. Kjellström, A. Poska, S. Wagner, M.-J. Gaillard, A.-K. Trondman, A. Mauri, B. A. S. Davis, J. O. Kaplan, H. J. B. Birks, A. E. Bjune, R. Fyfe, T. Giesecke, L. Kalnina, M. Kangur, W. O. van der Knaap, U. Kokfelt, P. Kuneš, M. Latał owa, L. Marquer, F. Mazier, A. B. Nielsen, B. Smith, H. Seppä, and S. Sugita. Regional climate model simulations for Europe at 6 and 0.2 k bp: sensitivity to changes in anthropogenic deforestation. *Clim. Past.*, 10(2):661–680, 2014.

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

S. Sugita, T. Parshall, R. Calcote, and K. Walker. Testing the landscape reconstruction algorithm for spatially explicit reconstruction of vegetation in northern Michigan and Wisconsin. *Quaternary. Res.*, 74(2):289–300, 2010.

H. Tjelmeland and K. V. Lund. Bayesian modelling of spatial compositional data. *J. Appl. Stat.*, 30(1):87–100, 2003.

A.-K. Trondman, M.-J. Gaillard, S. Sugita, R. Fyfe, J. Kaplan, A.-B. Nielsen, L. Marquer, F. Mazier, A. Poska, and G. Strandberg. Land cover-climate interactions in NW Europe, 6000 BP and 200 BP – first results of the Swedish LANDCLIM project. In *IPC/IOPC 2012; SS07-O14 (530)*, 2012.

A.-K. Trondman, M.-J. Gaillard, S. Sugita, F. Mazier, R. Fyfe, J. Lechterbeck, L. Marquer, A. Nielsen, C. Twiddle, P. Barratt, H. Birks, A. Bjune, C. Caseldine, R. David, J. Dodson, W. Dörfler, E. Fischer, T. Giesecke, T. Hultberg, M. Kangur, P. Kuneš, M. Latałowa, M. Leydet, M. Lindbaldh, F. Mitchell, B. Odgaard, S. Peglar, T. Persson, M. Rösch, P. van der Knaap, B. van Geel, A. Smith, and L. Wick. Pollen-based quantitative reconstructions of past landcover in NW Europe between 6k years BP and present for climate modelling. Submitted to Global Change Bio., 2014.

G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comp.*, 2(1):5–16, 1981.

D. Wårlind. *The Role of Carbon-Nitrogen Interactions for Terrestrial Ecosystem Dynamics under Global Change-a modelling perspective*. PhD thesis, Lund University, 2013.

P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

P. Whittle. Stochastic processes in several dimensions. *B. Int. Statist. Inst.*, 40(2): 975–994, 1963.

B

Paper B

Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties

Behnaz Pirzamanbein[1,2], Johan Lindström[1], Anneli Poska[3,4], Marie-José Gaillard[5]

[1] *Centre for Mathematical Sciences, Lund University, Sweden* [2] *Centre for Environmental and Climate Research, Lund University, Sweden* [3] *Department of Physical Geography and Ecosystems Analysis, Lund University, Sweden* [4] *Institute of Geology, Tallinn University of Technology, Estonia* [5] *Department of Biology and Environmental Sciences, Linnaeus University, Sweden*

Abstract

In this paper, we construct a hierarchical model for spatial compositional data, which is used to reconstruct past land-cover compositions (in terms of coniferous forest, broadleaved forest, and unforested/open land) for five time periods during the past 6 000 years over Europe . The model consists of a Gaussian Markov Random Field (GMRF) with Dirichlet observations. A block updated Markov chain Monte Carlo (MCMC), including an adaptive Metropolis adjusted Langevin step, is used to estimate model parameters. The sparse precision matrix in the GMRF provides computational advantages leading to a fast MCMC algorithm. Reconstructions are obtained by combining pollen-based estimates of vegetation cover at a limited number of locations with scenarios of past deforestation and output from a dynamic vegetation model. To evaluate uncertainties in the predictions a novel way of constructing joint confidence regions for the entire composition at each prediction location is proposed. The hierarchical model's ability to reconstruct past land cover is evaluated through cross validation for all time periods, and by comparing reconstructions for the recent past to a present day European

forest map. The evaluation results are promising and the model is able to capture known structures in past land-cover compositions.

Key words: Gaussian Markov Random Field, Dirichlet observation, Adaptive Metropolis adjusted Langevin, Pollen records, Confidence regions

1 Introduction

Modelling the spatial distribution in species composition and the relative abundances of different species is a common problem in environmental studies. (Aitchison, 1986, Billheimer et al., 2001, Paciorek and McLachlan, 2009, Pirzamanbein et al., 2014, Tjelmeland and Lund, 2003). In this paper we develop a statistical model for spatial compositional data and a way of assessing the uncertainties in the resulting compositional reconstructions at unobserved locations. The model is used to reconstruct past land-cover composition over Europe from local pollen-based estimates of vegetation cover.

1.1 Spatial Interpolation of Compositional Data

A common approach to modelling compositional data is Gaussian modelling of log-ratio transformed data (Aitchison, 1986), where the spatial structure can be captured using Gaussian fields (Billheimer et al., 2001, Pirzamanbein et al., 2014, Tjelmeland and Lund, 2003). However, modelling transformed compositions as Gaussian might understate the uncertainty in the data, especially in cases of overdispersion (Paciorek and McLachlan, 2009).

To capture the variability in our observations, we propose a Bayesian hierarchical model (described in Sec. 2) where the compositional data are seen as Dirichlet observations of an underlying latent field of probabilities. The field of compositional probabilities is in turn modelled using a transformed Gaussian Markov Random Field (GMRF) (Lindgren et al., 2011, Rue and Held, 2004). The sparsity in the precision matrix of the GMRF allows us to compute the Hessian for the entire latent field, allowing for fast estimation (see Sec. 3) using a Metropolis Adjusted Langevin algorithm (MALA) (Girolami and Calderhead, 2011, Roberts and Stramer, 2003).

To describe the uncertainties in the compositional reconstructions we propose a novel way of computing joint confidence and prediction regions for compositional data (Sec. 4). The method accounts for the interdependence among the

components of the compositional data and allows us to illustrate the joint uncertainty in the composition at each prediction location.

1.2 Climate Studies and Past Land Cover

For climate modelling studies the land-cover composition is commonly divided into three land cover types: coniferous forest, broadleaved forest, and unforested/open land. The spatial distribution of these land cover types play an important role in the climate system (Claussen et al., 2001). Accurate, spatially continuous, descriptions of past land cover types are necessary to assess past land cover-climate interactions (Brovkin et al., 2006) and the impact of anthropogenic land-cover changes on climate (Gaillard et al., 2010, 2015, Pirzamanbein et al., 2014, Strandberg et al., 2014).

Historic maps and surveys of past land cover have limited temporal coverage (rarely more than the past 300 to 500 years) and is often spatially fragmented due to a lack of transnational databases. Land-cover in climate models is currently implemented using a combination of dynamic vegetation models (e.g. LPJ-GUESS Smith et al., 2001) and scenarios of anthropogenic land-cover changes (e.g. Kaplan et al., 2009, Klein Goldewijk et al., 2011, Pongratz et al., 2008). Here, the dynamic vegetation models provide a climate-induced, potential vegetation, which is modified by the anthropogenic scenarios to account for human activities (mainly deforestation).

Land-cover reconstruction from fossil pollen records is an alternative, to the dynamic vegetation model simulations and anthropogenic scenarios, that may provide more realistic descriptions of past land cover for climate modelling studies (Gaillard et al., 2010, Trondman et al., 2015). Given pollen records extracted from lakes and bogs, pollen-based estimates of vegetation cover are obtained using a model (here the REVEALS model of Sugita, 2007a,b). The model provides estimates of pollen-based land-cover composition (hereafter called PbLCC) for a limited area (ca. 100 km x 100 km) around each lake or bog. For use in climate modelling these PbLCC estimates need to be interpolated into continuous maps of past land-cover composition at sub-continental to global scales (Paciorek and McLachlan, 2009, Pirzamanbein et al., 2014).

The PbLCC data used in this paper are available for five time periods during the past 6 000 years, and the proposed Bayesian model and estimation procedure is used to interpolate the PbLCC data for each time period. The results are validated using present-time forest maps and cross-validation (Sec. 5.3). The model

shows good predictive power, capturing known structures and historical changes in land-cover composition.

The paper ends with some brief conclusion in Sec. 6.

2 Model

To model the spatial structure in the compositional data we propose a hierarchical model, where the observed compositions at each location are modelled as draws from a Dirichlet distribution. The Dirichlet is parametrized using a scale (or concentration) parameter and a vector of probabilities. The spatial dependence in these compositional probabilities is modelled using a transformed GMRF. Details regarding the observational model are given in Sec. 2.1, and Sec. 2.2 describes the latent field.

2.1 Dirichlet Distribution and Link Function

Compositional data are discussed in detail by Aitchison (1986), here a brief overview is given. Let $\mathbf{y}_s = (y_{s,1}, y_{s,2}, \cdots, y_{s,D})$ be the D-compositional data at location $u_s \in \mathbb{R}^2,\ s = 1, \cdots, N_o$, the restrictions for compositional data imply that: $y_{s,k} \in (0,1)$ and $\sum_{k=1}^{D} y_{s,k} = 1$. Conditional on the transformed underlying field, $\mathbf{z} = f(\boldsymbol{\eta})$, we assume that the data, $\mathbf{Y} = \{\mathbf{y_s}\}_{s=1}^{N_o}$, are independent draws from a multivariate Dirichlet distribution,

$$\mathbb{P}(\mathbf{Y}|\alpha, \mathbf{z}) = \prod_{s=1}^{N_o} \Big(\frac{\Gamma(\alpha)}{\prod_{k=1}^{D} \Gamma(\alpha z_{s,k})} \prod_{k=1}^{D} y_{s,k}^{\alpha z_{s,k}-1} \Big), \quad \alpha > 0, \tag{1}$$

and

$$z_{s,k} \in (0,1), \qquad \sum_{k=1}^{D} z_{s,k} = 1$$

where α is a Dirichlet scale parameter.

The link function, f, between $\mathbf{z}$ and $\boldsymbol{\eta}$ can be any function from $\mathbb{R}^{d \times N_o}$ to $(0,1)^{D \times N_o}$ such that:

$$\begin{aligned} f(\eta_1, \cdots, \eta_d) &= (Z_1, \cdots, Z_d, Z_D), \\ \sum_{k=1}^{D} Z_k &= 1, \quad \text{and} \quad d = D - 1. \end{aligned} \tag{2}$$

Here Z_k is a $N_o \times 1$ column vector containing the k^{th} component of the D-compositional data and η_k is a column vector with the k^{th} latent field, i.e. the probabilities and latent fields for location s are given by $\{z_{s,k}\}_{k=1}^{D}$ and $\{\eta_{s,k}\}_{k=1}^{d}$, respectively.

In this paper the link function is constructed by applying the additive log-ratio (Aitchison, 1986) transform

$$\eta_{s,k} = \log z_{s,k} - \log z_{s,D}, \quad k = 1, \ldots, D-1 \tag{3}$$

for each location s. Possible choices for transforms at each location also include the isometric log-ratio transformation (Egozcue et al., 2003). However, it excludes the central log-ratio transformation (Aitchison, 1986), which gives a latent $\boldsymbol{\eta}$-field with unidentifiable mean.

2.2 Latent Field

Given a total of $N \geq N_o$ locations at which we want to provide composition predictions the latent field, $\boldsymbol{\eta}_{all}$, is multivariate with $d = D - 1$ elements at each location ($N \geq N_o$ since we are providing predictions at the observed and additional locations). To simplify notation the latent field is represented as a $Nd \times 1$ vector $\boldsymbol{\eta}_{all} = (\eta_{all,1}^{\top}, \cdots, \eta_{all,d}^{\top})^{\top}$, where each $\eta_{all,k}$ is spatial field with N locations.

The latent field and its connection to the observed locations is given as:

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{A}\boldsymbol{\eta}_{all} \\ \boldsymbol{\eta}_{all} &= \mathbf{B}\boldsymbol{\beta} + \mathbf{X}. \end{aligned} \tag{4}$$

where $\mathbf{A} = \mathbb{I}_{d\times d} \otimes A$ extracts the observed elements from $\boldsymbol{\eta}_{all}$, with A being a $N_o \times N$ sparse observation matrix; $\mathbf{B} = \mathbb{I}_{d\times d} \otimes B$ with B being a $N \times p$ matrix of covariates; $\boldsymbol{\beta}$ is a $dp \times 1$ matrix of regression coefficients; and $\mathbf{X} = (X_1^{\top}, \cdots, X_d^{\top})^{\top}$ is a spatially correlated multivariate field. With this structure, the spatial dependence $\mathbf{X}$, can be modelled as a GMRF with a separable covariance structure, i.e. $\boldsymbol{\rho} \otimes \mathbf{Q}^{-1}$, which captures the dependency among and within the fields;

$$\mathbf{X} \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{\rho} \otimes \mathbf{Q}^{-1}(\kappa)\right). \tag{5}$$

Here $\boldsymbol{\rho}$ is a $d \times d$ matrix of covariances among the d multivariate fields (X_k, $k = 1, \cdots, d$), and $\mathbf{Q}(\kappa)$ is a $N \times N$ precision matrix of a GMRF with spatial scale

parameter κ. $\mathbf{Q}$ is chosen as a precision matrix which approximates a stationary Matérn field (Matérn, 1960) with smoothness $\nu = 1$;

$$\mathbf{Q}(\kappa) = \kappa^4\mathbf{C} + 2\kappa^2\mathbf{G} + \mathbf{G}\mathbf{C}^{-1}\mathbf{G}. \tag{6}$$

Here $\mathbf{C}$ is a diagonal matrix and $\mathbf{G}$ is a finite difference approximation of the negative Laplacian (cf. Appendix A in Lindgren et al., 2011). This precision is also a solution to a stationary stochastic partial deferential equation (SPDE) field with $\alpha = 2$ (see Lindgren et al., 2011, for details). While the smoothness, ν, of the latent field is known to affect spatial prediction (Stein, 1999) it is also very hard to estimate (Haran, 2011) and a popular default is to use the exponential covariance ($\nu = 0.5$). For GMRF models in $\mathbb{R}^2$, ν has to be integer resulting in our choice of $\nu = 1$; a value also suggested by Whittle (1954). It should further be noted that for $\nu = 1$ (or $\alpha = 2$) the special-case of $\kappa = 0$ (infinite range) gives Wahba (1981) splines, providing a link between spline smoothing and Gaussian spatial-processes (Kimeldorf and Wahba, 1970, Nychka, 2000).

2.3 Hierarchical Model and Priors

The full hierarchical model (Fig. 1) based on Dirichlet observations (1) of a transformed latent GMRF (5) becomes

$$\begin{aligned}
\mathbf{y_s}|\alpha,\boldsymbol{\eta} &\sim \mathsf{Dir}\left(\alpha, f_s(\boldsymbol{\eta})\right), & s &= 1,\ldots,N_o \\
\boldsymbol{\eta}_{all} &= \mathbf{B}\boldsymbol{\beta} + \mathbf{X}, & \boldsymbol{\eta} &= \mathbf{A}\boldsymbol{\eta}_{all}, \\
\mathbf{X}|\kappa,\boldsymbol{\rho} &\sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{\rho}\otimes\mathbf{Q}^{-1}(\kappa)\right), & & \\
\boldsymbol{\beta} &\sim \mathsf{N}\left(0, \mathbb{I}q_\beta^{-1}\right), & \alpha &\sim \Gamma\left(a_\alpha, b_\alpha\right) \\
\kappa &\sim \Gamma\left(a_\kappa, b_\kappa\right), & \boldsymbol{\rho}|\kappa &\sim \mathsf{IW}\left(a_\rho\mathbb{I}, b_\rho\right)
\end{aligned} \tag{7}$$

where $\mathbb{I}$ are appropriate identity matrices. To make $\mathbf{X}$ and $\boldsymbol{\beta}$ jointly normal, we use a vague Gaussian prior for $\boldsymbol{\beta}$ with precision $q_\beta = 10^{-3}$. The Dirichlet scale parameter, α, and spatial scale parameter, κ, are given gamma priors, and for $\boldsymbol{\rho}$ we choose a conjugate prior for covariance matrices, the inverse Wishart (*IW*). The conjugacy of the inverse Wishart provides computational advantages when updating the parameters of our multivariate latent field.

A suitable prior on κ can be obtained by noting its link with the range of the field: range $\approx \sqrt{8\nu}/\kappa$ (Lindgren et al., 2011). This link is used by R-INLA to create reasonable defualt priors where the range is related to the size of the domain

(Lindgren and Rue, 2015). An alternative option is presented by Fuglstad et al. (2016) and Simpson et al. (2015), introducing a prior that shrinks towards $\kappa = 0$, the intrinsic field (i.e. a spline smoother). This prior is motivated by the intrinsic field representing a simpler model, to be prefered in the absence of convincing data. The prior in Fuglstad et al. (2016) corresponds to $a_\kappa = 1$ with b_κ chosen to give a suitably small prior-probability of short ranges. A 1% probability of range < 1 results in $b_\kappa = -\log(0.01) \cdot 1/\sqrt{8}$, the range of 1 is based on the unit distance between our gridcell centroids. For the inverse Wishart prior on $\boldsymbol{\rho}$, we chose an uninformative prior with $a_\rho = 1$ and $b_\rho = 10$. The inverse Wishart is proper if the degree of freedom is $b_\rho > d - 1$, has finite mean if $b_\rho > d + 1$ and has finite variance if $b_\rho > d + 3$. In practice b_ρ is often chosen somewhat larger than these lower bounds (see e.g. Schmidt et al., 2010). Given our lack of intuition for α we pick uninformative prior resulting in the following values of the hyper-parameters:

$$
\begin{aligned}
a_\alpha &= 1.5, & a_\kappa &= 1, & a_\rho &= 1, \\
b_\alpha &= 0.1, & b_\kappa &= \frac{\log(100)}{\sqrt{8}}, & b_\rho &= 10.
\end{aligned}
$$

Having detailed the model, parameter estimation and reconstruction of $\boldsymbol{\eta}_{\mathbf{all}}$, using MCMC, are described in the following section.

3 Estimation Using MCMC

A block-updated MCMC algorithm is used to estimate the latent field $\boldsymbol{\eta}_{all}$ and the unknown parameters $\alpha, \kappa, \boldsymbol{\rho}$. For GMRFs, joint updating of parameters in as large blocks as possible has been shown to improve mixing and convergence (Knorr-Held and Rue, 2002). Therefore, the algorithm in this paper updates the unknowns by alternating between two blocks: the first block updates the latent fields and the Dirichlet scale parameter using MALA (Girolami and Calderhead, 2011, Roberts and Stramer, 2003), the second block updates the parameters of the GMRF, κ and $\boldsymbol{\rho}$, using a combination of random walk proposals and the conjugate posterior for $\boldsymbol{\rho}$.

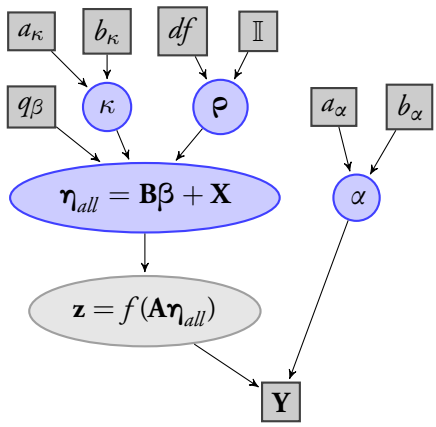


Figure 1: Directed acyclic graph describing the conditional dependencies in the hierarchical model.

3.1 Updating $\boldsymbol{\eta}_{all}$ and α

To update $\boldsymbol{\eta}_{all} = \mathbf{B}\beta + \mathbf{X}$ and α we use a Metropolis-Hastings step to draw samples from the conditional distribution

$$\begin{aligned}\mathbb{P}(\mathbf{X}, \beta, \alpha|\kappa, \rho, \mathbf{Y}) &\propto \Big(\prod_{s=1}^{N_o} \mathbb{P}(\mathbf{y}_s|f_s\big(\mathbf{A}\boldsymbol{\eta}_{all}\big), \alpha\big) \cdot \mathbb{P}(\mathbf{X}|\kappa, \rho) \cdot \mathbb{P}(\beta) \cdot \mathbb{P}(\alpha) \\ &\propto \prod_{s=1}^{N_o} \Big(\frac{\Gamma(\alpha)}{\prod_{k=1}^{D}\Gamma(\alpha z_{s,k})} \prod_{k=1}^{D} y_{s,k}^{\alpha z_{s,k}-1}\Big) \\ &\quad \cdot \exp\Big(-\frac{1}{2}\mathbf{X}^\top\big(\rho^{-1} \otimes \mathbf{Q}(\kappa)\big)\mathbf{X}\Big) \\ &\quad \cdot \exp\Big(-\frac{q_\beta}{2}\beta^\top\beta\Big) \cdot \alpha^{a_\alpha-1}e^{-\alpha \cdot b_\alpha}.\end{aligned} \tag{8}$$

The Metropolis-Hastings step uses a MALA proposal:

$$\mathbf{X}^*, \beta^*, \alpha^*|\mathbf{X}, \beta, \alpha \quad \sim \quad \mathsf{N}\left((\mathbf{X}, \beta, \alpha)^\top + \frac{\varepsilon^2}{2}\boldsymbol{\mathcal{I}}^{-1}\nabla l, \varepsilon^2\,\boldsymbol{\mathcal{I}}^{-1}\right), \tag{9}$$

where ε is the step size of MALA, ∇l is a vector of derivatives of $\log \mathbb{P}(\mathbf{X}, \beta, \alpha|\kappa, \rho, \mathbf{Y})$ w.r.t. $\mathbf{X}, \beta$ and α (for computational details see Appendix A.2) and $\boldsymbol{\mathcal{I}}$ is the expected Fisher information matrix, (see Appendix A.3). At each iteration, $\boldsymbol{\mathcal{I}}^{-1}\nabla l$

gives a sampling direction from the current state which is similar to a Newton-Raphson step (Givens and Hoeting, 2012, Ch. 2). Further, the proposal variance, $\boldsymbol{\mathcal{I}}^{-1}$, accounts for the dependency among the parameters. Due to the GMRF structure of the latent fields, $\boldsymbol{\mathcal{I}}$ will be a sparse matrix reducing the computations to sampling from a GMRF, for which efficient algorithms exist (Rue and Held, 2004).

In order to get reasonable acceptance rate, an adaptive MCMC method (Andrieu and Thoms, 2008) is used for ε with the following updating rule;

$$\varepsilon_{i+1} = \varepsilon_i + \gamma_{i+1}(\widehat{acc}_{\mathbf{X},\boldsymbol{\beta},\alpha}(\varepsilon_i) - 0.57) \tag{10}$$

where ε_i is the step size for the i^{th} MCMC iteration, $\gamma_i = i^{-1/2}$, $\widehat{acc}$ is the acceptance probability of the ith step, and 0.57 is the target acceptance rate for a MALA proposal as suggested by Roberts and Rosenthal (1998).

3.2 Updating κ and $\boldsymbol{\rho}$

The second block is updated using a combination of the conjugate posterior for $[\boldsymbol{\rho}|\mathbf{X}, \kappa]$ and a Metropolis-Hastings random walk (in log scale) for $[\kappa|\mathbf{X}]$. The joint posterior of $[\kappa, \boldsymbol{\rho}|\mathbf{X}]$ can be written as

$$\mathbb{P}(\kappa, \boldsymbol{\rho}|\mathbf{X}) = \mathbb{P}(\boldsymbol{\rho}|\mathbf{X}, \kappa) \cdot \mathbb{P}(\kappa|\mathbf{X}). \tag{11}$$

Due to the conjugate prior for $\boldsymbol{\rho}$ the conditional posterior for $\boldsymbol{\rho}$ is inverse Wishart;

$$[\boldsymbol{\rho}|\kappa, \mathbf{X}] \propto \mathsf{IW}\left(a_\rho \mathbb{I} + \mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}, N + b_\rho\right) \tag{12}$$

with $\mathbf{x}$ being a $N \times d$ matrix given by $\mathbf{x} = [X_1, \cdots, X_d]$. The conjugacy makes it possible to marginalize over $\boldsymbol{\rho}$ (see Appendix B) giving

$$\mathbb{P}(\kappa|\mathbf{X}) \propto \int \mathbb{P}(\boldsymbol{\rho}|\kappa, \mathbf{X})\mathbb{P}(\kappa) d\boldsymbol{\rho} \propto \frac{a_\rho^{\frac{db_\rho}{2}} |\mathbf{Q}(\kappa)|^{\frac{d}{2}}}{|a_\rho \mathbb{I} + \mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}|^{\frac{N+b_\rho}{2}}} \mathbb{P}(\kappa). \tag{13}$$

Samples from $[\kappa, \boldsymbol{\rho}|\mathbf{X}]$ are now obtained by first sampling from the posterior (13) using a Metropolis-Hastings step with a random-walk proposal in log scale,

$$\log \kappa^* = \log \kappa + \varepsilon_\kappa, \qquad \varepsilon_\kappa \sim \mathsf{N}\left(0, \sigma_\kappa^2\right).$$

Given a proposal κ^*, $\boldsymbol{\rho}^*$ is sampled from (12). These two steps can be seen as a joint Metropolis-Hastings step for $\boldsymbol{\rho}$ and κ with proposal density $q(\kappa^*, \boldsymbol{\rho}^*|\kappa, , \boldsymbol{\rho}) = \mathbb{P}(\boldsymbol{\rho}^*|\mathbf{X}, \kappa^*) \cdot q(\kappa^*|\kappa)$ and acceptance ratio:

$$\begin{aligned}
\text{acc}_{\kappa,\boldsymbol{\rho}} &= \min\left(1, \frac{\mathbb{P}(\kappa^*, \boldsymbol{\rho}^*|\mathbf{X})}{\mathbb{P}(\kappa, \boldsymbol{\rho}|\mathbf{X})} \cdot \frac{q(\kappa, \boldsymbol{\rho}|\kappa^*, \boldsymbol{\rho}^*)}{q(\kappa^*, \boldsymbol{\rho}^*|\kappa, \boldsymbol{\rho})}\right) \\
&= \min\left(1, \frac{\mathbb{P}(\boldsymbol{\rho}^*|\kappa^*, \mathbf{X}) \cdot \mathbb{P}(\kappa^*|\mathbf{X})}{\mathbb{P}(\boldsymbol{\rho}|\kappa, \mathbf{X}) \cdot \mathbb{P}(\kappa|\mathbf{X})} \cdot \frac{\mathbb{P}(\boldsymbol{\rho}|\mathbf{X}, \kappa) \cdot q(\kappa|\kappa^*)}{\mathbb{P}(\boldsymbol{\rho}^*|\mathbf{X}, \kappa^*) \cdot q(\kappa^*|\kappa)}\right) \\
&= \min\left(1, \frac{\mathbb{P}(\kappa^*|\mathbf{X})}{\mathbb{P}(\kappa|\mathbf{X})} \cdot \frac{\kappa^*}{\kappa}\right).
\end{aligned} \tag{14}$$

Since the acceptance ratio depends only on κ we can delay the sampling of $\boldsymbol{\rho}^*$ until we know if the suggested κ^* has been accepted.

The proposal variance, σ_κ^2, is determined using an adaptive scheme similar to (10), with target acceptance rate of 0.44 (Roberts et al., 1997). The difference in target acceptance rate is due to the difference between MALA and random-walk Metropolis-Hastings (see Rosenthal, 2011, for a discussion).

4 Uncertainty

To obtain uncertainties in the composition estimates at each location, we use the MCMC samples of $\boldsymbol{\eta}$ at each location. Given the model structure with a Gaussian prior for $\boldsymbol{\eta}$ we base the joint confidence regions for the composition estimates on the elliptical confidence regions obtained for multivariate Gaussian distributions. Using the sample mean, $\boldsymbol{\mu}$, and the sample covariance, $\boldsymbol{\Sigma}$, in the MCMC samples, we construct the confidence region for each location as the ellipse

$$(\boldsymbol{\eta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\eta} - \boldsymbol{\mu}) = C_\alpha. \tag{15}$$

The quantile C_α is taken as the α-quantile of the above squared Mahalanobis distance computed for all the MCMC samples (for a multivariate Gaussian $C_\alpha = \chi^2_\alpha(d)$). Thereafter, the confidence ellipse is transformed from R^d to $(0,1)^D$ using (2). For illustration purposes, we choose $D = 3$. The new ternary region is considered as 95% confidence region for the transformed $\boldsymbol{\eta}$, i.e. the composition estimates, see Fig. 2.

To illustrate the changes in compositions, we choose the maximum and minimum along each dimensions of the ternary plot, i.e. in each component. This

way, we get a joint lower bound (minimum) and upper bound (maximum) for each composition together with the corresponding changes in the other compositions "most likely" to occur at the bounds (Fig. 2).

In addition, we compute prediction regions for the compositions. To obtain prediction regions, we simulate new Dirichlet observations for each MCMC sample of $\boldsymbol{\eta}$ and α. These D-composition Dirichlet-simulations are then transformed to $\mathbb{R}^d$ using the link function. The procedure above is then used to obtain prediction ellipses and ternary prediction regions.

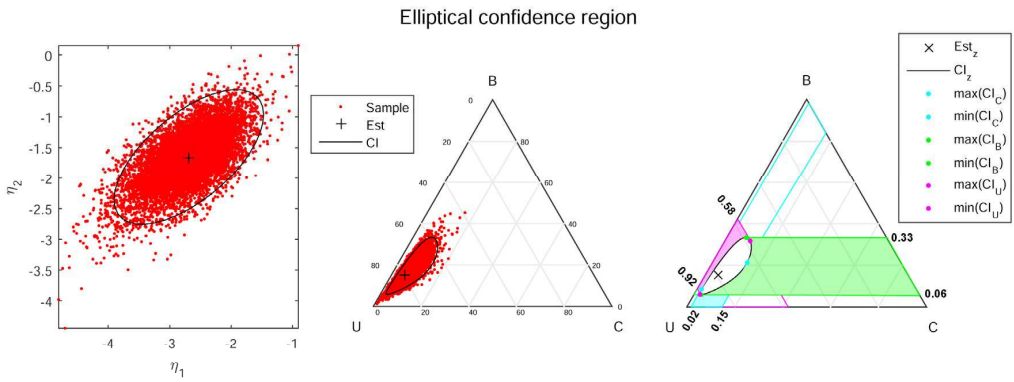


Figure 2: The left plot shows the 95% elliptical confidence region for the $\boldsymbol{\eta}$ samples at location s. The middle ternary diagram shows the transformed samples and ellipse. The right hand ternary diagram shows the joint maximum and minimum in each composition, C, B, and U; together with the confidence interval for the other two compositions.

5 Application

The model presented in Sec. 2 was applied to the PbLCC data with the goal of reconstructing past land cover over Europe. Two versions of the model in (4) were considered: 1) a full spatial model with $\boldsymbol{\eta}_{all} = \mathbf{B}\boldsymbol{\beta} + \mathbf{X}$ (includes all parameters, $\alpha, \boldsymbol{\beta}, \boldsymbol{\rho}, \kappa$ and the $\mathbf{X}$ fields), and 2) a regression model with no spatial dependence structure where $\boldsymbol{\eta}_{all} = \mathbf{B}\boldsymbol{\beta}$; the regression model is included to allow an evaluation of the need for spatial structure. The intrinsic GMRF model ($\kappa = 0$), also considered in Pirzamanbein et al. (2014), performed similar to or slightly worse

than the full spatial model with Dirichlet observations, hence those results have been excluded for brevity.

The remainder of this section consists of: a description of the data (Sec. 5.1), parameters estimation and spatial reconstruction results for the models (Sec. 5.2), and validation of model performance (Sec. 5.3).

5.1 Data

The data used for the reconstruction of past land-cover composition over Europe consists of pollen-based REVEALS estimates (here called pollen-based land-cover composition data — PbLCC) of the three land cover types: Coniferous forest, Broadleaved forest and Unforested land. The PbLCC data was obtained using the REVEALS model (Sugita, 2007b) and a detailed description of the data is given in (Trondman et al., 2015). REVEALS is mechanistic model that takes into account the size of sedimentary basins and inter-taxonomic differences in pollen productivity and dispersal to estimate regional vegetation cover from pollen records. Trondman et al. applied REVEALS to 636 pollen records from lakes and bogs; producing estimates of regional land cover for 25 plant taxa which were then grouped into the 3 land cover types (see Table 1 and Appendix S2 in Trondman et al., 2015, for details). The regional estimates from REVEALS were obtained for $1^\circ \times 1^\circ$ grid cells, with each estimate being based on the pollen records from all lakes and bogs within that grid cell (Hellman et al., 2008, Trondman et al., 2015). The $1^\circ \times 1^\circ$ grid is an appropriate scale for climate models, which currently work at this or higher resolutions (Trondman et al., 2015). Since the sedimentary pollen records used by REVEALS are obtained from lakes and bogs the grid based REVEALS estimates are limited to providing land cover in grid cells surrounding the lakes and bogs. This leads to a PbLCC dataset with incomplete coverage across Europe, which needs to be interpolated to produce land cover compositions for the entire region.

The PbLCC data are available for five time periods centred around 1900, 1725 and 1425 CE, 1000 and 4000 BCE, with 175, 181, 193, 204 and 196 observed grid cells, respectively (Trondman et al., 2015); strictly the time periods are: present–1850 CE, 1850–1600 CE, 1600–1250 CE, 1250–750 BCE, and 4250–3750 BCE; where "present" should be interpreted as the most recent pollen records recovered at each site. These time periods are commonly used in both climate modelling and palaeoecological studies since they represent major climatic and historical events; Recent Past, Little Ice Age, Black Death, Late Bronze Age,

and Early Neolithic.

To capture large scale structures in the land-cover composition, covariates consisting of potential natural vegetation cover adjusted for human land use, and elevation were used. The choice of covariates was based on the best model found in Pirzamanbein et al. (2014), and detailed descriptions of the covariates can be found in that paper. Here we only provide a brief summary.

Dynamic vegetation model based estimates of climate-induced potential natural vegetation, for the study area and specified time periods, were obtained using the LPJ-GUESS model (Smith et al., 2001). To account for human land use, the potential natural vegetation was adjusted for anthropogenic deforestation using the KK10 scenarios of Kaplan et al. (2009). The KK10 scenarios provide assessments of human induced deforestation based on estimates of past human population densities, land area required for food production to sustain that population, and a model of land suitability for food production. Combining the potential natural vegetation cover from LPJ-GUESS and the KK10 scenarios of deforestation resulted in a land cover covariate, denoted LPJ-GUESS$_{\text{KK}}$.

The elevation data were obtained from the Shuttle Radar Topography Mission (Becker et al., 2009)[1] and upscaled by averaging from the original resolution of 3 arc-seconds to the $1^\circ \times 1^\circ$ grid cells. The upscaled data was truncated to ≥ 0 to handle a few grid cells along the Norwegian coast which otherwise would have negative average elevation due to the presence of deep coastal fjords.

Since the potential land cover, LPJ-GUESS$_{\text{KK}}$, is compositional it was transformed using (2), and the covariate matrix, $\mathbf{B}$ consisted of the following columns: B_0 – intercept; B_1, B_2 – additive log ratio transformed LPJ-GUESS$_{\text{KK}_{1,2}}$; and B_3 – elevation.

To evaluate our results we used present-time European forest maps compiled by the European Forest Institute. These maps are based on a combination of satellite data (NOAA-AVHRR) and national forest-inventory statistics from 1990–2005 (Päivinen et al., 2001, Schuck et al., 2002) [2]. The European Forest Institute forest maps (EFI-FM; with proportions of coniferous- and broadleaved-forest cover) were upscaled by averaging from 1 km $\times$1 km to $1^\circ \times 1^\circ$ resolution. The proportions of unforested area were calculated by subtracting the total sum of

[1] downloaded from `ftp://topex.ucsd.edu/pub/srtm30_plus/` on $2011-09-03$

[2] downloaded from the European Forest Institute webpage `http://www.efi.int/portal/virtual_library/information_services/mapping_services/forest_map_of_europe`

forested cover from 1.

5.2 Results

To estimate the parameters for each model, we ran 100 000 MCMC iterations with a burn-in sample size of 10 000. Diagnostics for the chains indicate a fast convergence for α, ρ and β; autocorrelation plots show good mixing of all parameters after burn-in.

Parameter estimates for the 1900 CE time period are given in Table 1; the parameter estimations for the other time periods can be found in Appendix C. Note that the α estimate for the regression model is lower than for the full spatial model, indicating higher observational variation in regression model.

Table 1: Parameter estimates (Est) and 95% confidence intervals (CI) for the two models (Full — spatial model and RM — regression model) fitted to the PbLCC data from the 1900 CE time period.

1900 CE

	Full		RM	
Parameter	Est	(CI)	Est	(CI)
α	10.86	(8.10 , 15.41)	6.36	(5.58 , 7.18)
κ	0.28	(0.14 , 0.45)	-	-
ρ_{11}	0.78	(0.12 , 2.61)	-	-
ρ_{12}	0.57	(0.05 , 1.96)	-	-
ρ_{22}	0.60	(0.10 , 1.97)	-	-
β_{10}	-0.68	(-1.64 , 0.15)	-0.13	(-0.25 , -0.02)
β_{11}	0.16	(0.08 , 0.24)	0.24	(0.22 , 0.27)
β_{12}	0.02	(-0.09 , 0.14)	-0.03	(-0.09 , 0.02)
β_{13}	0.05	(-0.15 , 0.26)	-0.10	(-0.19 , -0.01)
β_{20}	-0.94	(-1.83 , -0.22)	-0.38	(-0.51 , -0.26)
β_{21}	0.04	(-0.05 , 0.12)	0.13	(0.11 , 0.16)
β_{22}	0.01	(-0.09 , 0.11)	-0.05	(-0.10 , 0.00)
β_{23}	-0.04	(-0.24 , 0.16)	-0.24	(-0.34 , -0.14)

Reconstructions of the land-cover composition for the two models and the 1900 CE time period are shown in Fig. 3. Results for the other time periods are available in Appendix D. Figure 3 shows that the land-cover reconstructions from the two models captured the structure in the PbLCC data. However, the results

from regression model is smoother than from the Full model. The Full model better captures the high abundance of unforested land in Poland, Denmark and south east Norway.

The uncertainties in the land-cover reconstructions were computed using the method described in Sec. 4. Results for the 1900 CE time period are presented in Fig. 4 and 5, with results for the remaining time periods given in Appendix E. The confidence and prediction regions represent the uncertainty in the latent field reconstruction, $\mathbf{z_s}$ and the potential uncertainty in new PbLCC data, $\mathbf{y_s}$, for a given grid cell, respectively. In general the full spatial model has larger confidence regions but smaller prediction regions than the regression model (Fig. 4). This is due to the spatial component in the Full model being able to better capture spatial variation resulting in a lower uncertainty (larger α) in the Dirichlet observations as compared to regression model. The maps of confidence regions (Fig. 5) illustrate rather large uncertainties in the predicted land-cover composition in general, and especially for Southeast Europe, a region with very few observations.

5.3 Validation

To evaluate the performance of the models, we compared the land-cover reconstructions for 1900 CE to the EFI-FM by computing the average compositional distances (ACD). The compositional distances (Aitchison, 1986, 1992, Aitchison et al., 2000) were computed for each location, using

$$\mathrm{ACD}(\mathbf{u},\mathbf{v}) = [(\mathbf{u}-\mathbf{v})^T\mathbf{J}^{-1}(\mathbf{u}-\mathbf{v})]^{1/2} \tag{16}$$

where $\mathbf{u}$ and $\mathbf{v}$ are additive log-ratio transforms of the compositions to be compared and $\mathbf{J}$ is a $d \times d$-matrix with elements $J_{p,l} = 2$ if $p = l$, and $J_{p,l} = 1$ if $p \neq l$. These compositional distances are then averaged over all grid cells. In terms of the original compositions, $\mathbf{p^u}$ and $\mathbf{p^v}$, the distance in (16) can be written as (Aitchison et al., 2000)

$$\mathrm{ACD}(\mathbf{p^u},\mathbf{p^v}) = \left[\sum_{i=1}^{D}\left(\log\frac{p_i^u}{g(\mathbf{p^u})} - \log\frac{p_i^v}{g(\mathbf{p^v})}\right)\right]^{1/2},$$

where $g(\mathbf{p})$ is the geometric mean, $g(\mathbf{p}) = \sqrt[D]{p_1 p_2 \cdots p_D}$.

Although a temporal misalignment exists between the PbLCC data (PbLCC data are from 1850 to the present) and the EFI-FM (inventory and satellite data

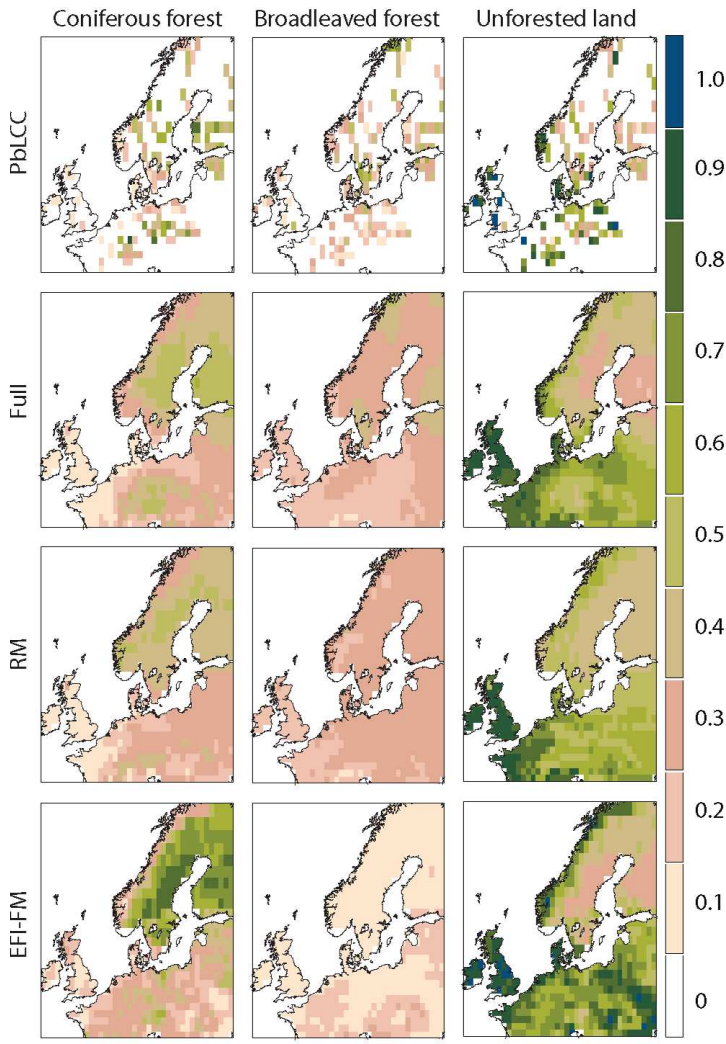


Figure 3: Results for the 1900 CE time period: the top row shows the PbLCC data from REVEALS, the bottom row shows the EFI-FM and the remaining rows show the reconstructions for the full spatial model (Full) and the regression model (RM). For larger maps see E.1.

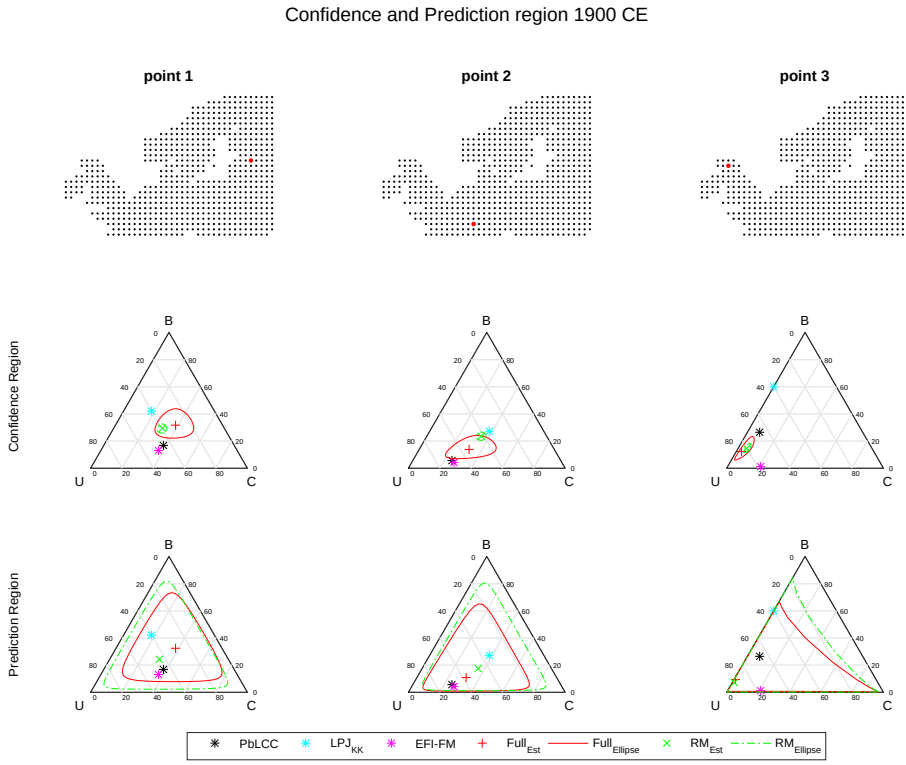


Figure 4: The first row shows the locations of the three selected grid cells. The second row shows the ternary confidence regions and the land-cover reconstructions for the two models (Full — spatial model and RM — regression model) together with the PbLCC data from REVEALS, the LPJ-GUESS$_{KK}$ land cover covariate and the EFI-FM for each location. The third row shows the ternary prediction regions.

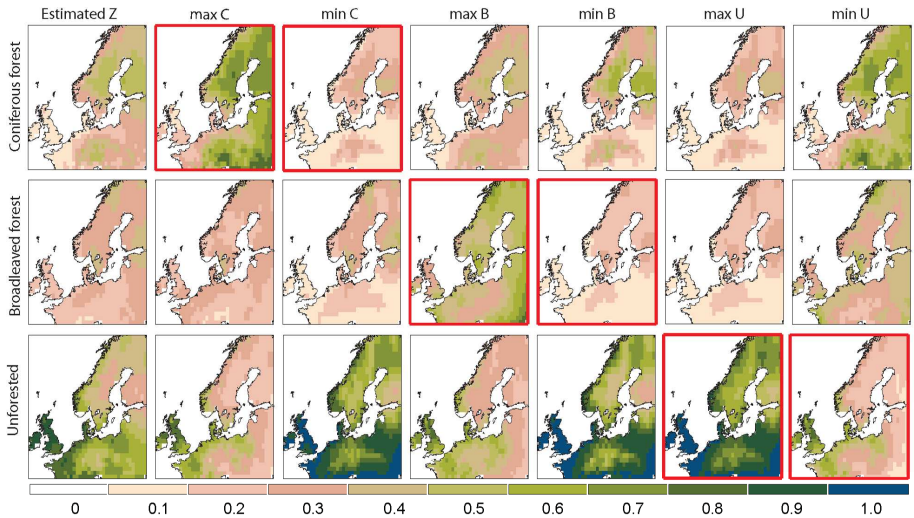


Figure 5: The first column shows the reconstructed land-cover composition for the 1900 CE time period using the full spatial model. Columns 2 and 3, row 1 (with thick/red axes), show the maximum and minimum of 95% elliptical confidence regions for Coniferous; rows 2 and 3 give the corresponding Broadleaved and Unforested compositions. Columns 4 and 5 (row 2 with thick/red axes) gives the bounds for the Broadleaved composition while columns 6 and 7 show the bounds for Unforested land (row 3 with thick/red axes). The concept of joint confidence interval for compositions is illustrated in Fig 2.

are from 1990-2005); EFI-FM provides the best complete and consistent land cover map of Europe for present times, making it a reasonable choice for the comparison. Figure 3 shows the maps of PbLCC data and EFI-FM. The main differences between the EFI-FM data and the PbLCC data for the 1900 CE time period are: 1) a lower abundance of broadleaved forest around most of Europe, 2) a higher abundance of coniferous forest in Sweden and Finland, and 3) a higher abundance of unforested land in North Norway in the EFI-FM data than in the PbLCC data. Compositional distances between land-cover reconstructions and EFI-FM were computed using (16) and averaged over all grid cells. The resulting ACD are 1.4757 and 1.5025 for the full spatial model and the regression model, respectively. This indicates that the full spatial model provides a reconstruction closer to EFI-FM than the regression model.

These results can also be compared to a model with Gaussian observations of transformed latent fields, (Pirzamanbein et al., 2014). The resulting ACD of the Gaussian observation models compare to the EFI-FM are 1.6007 (for the intrinsic GMRF model) and 1.6140 for the regression model. The differences between what Pirzamanbein et al. (2014) reported (1.5201 and 1.5177, respectively) and our results using their models are due to an increase in available data leading to more grid cells in our reconstructions. These results indicate smaller distances between the land-cover reconstructions and EFI-FM for the models with Dirichlet observations proposed in this paper compared to similar models with Gaussian observations.

Since no ground truth exists for the other time periods, we applied a 6-fold cross-validation scheme for the models for each of the five time periods (Friedman et al., 2001, Ch. 7.10). The cross-validation was run for 10 different, randomly selected 6 folds to assess the variability due to different cross validation groupings. Average compositional errors and standard deviations are shown in Table 2. The full spatial model gives the best predictions for all the five time periods.

6 Conclusion

In this paper we have introduced a model for spatial interpolation of compositional data that relies on Dirichlet observations of an underlying multivariate GMRF. In theory the formulation allows for a wide class of link-functions between the GMRF and the compositional probabilities in the Dirichlet observations; we used the additive log ratio transformation throughout the paper. Since the sparse

Table 2: Average compositional error (and standard deviation) from 10 different 6-fold cross-validations for each of the models, and time periods.

	Full		Regression	
Time	CV_{error}	(sd)	CV_{error}	(sd)
1900 CE	**1.0169**	(0.0122)	1.1439	(0.0061)
1700 CE	**1.1448**	(0.0084)	1.2891	(0.0054)
1400 CE	**1.2009**	(0.0071)	1.4061	(0.0042)
1000 BCE	**1.3260**	(0.0083)	1.5287	(0.0062)
4000 BCE	**1.2131**	(0.0109)	1.3396	(0.0045)

structure in the precision matrix of the GMRF carries over to the expected Fisher information used in MALA, the model formulation with a latent GMRF allows for fast MCMC-based estimation of parameters and latent field. As a result our MCMC produced 10 samples per second using MATLAB® on a standard desktop (Intel® Core™ *i7* − 2600 CPU (2011) with 8 GB memory) for a latent field with 2160 nodes (bivariate field on a 27-by-40 grid); resulting in a total run time of less than 3 hours for a chain with 100 000 samples.

To evaluate prediction uncertainties we also proposed a method for construction of joint confidence and prediction regions of the predicted compositions at each location. The idea behind the method is to use the MCMC samples to first construct elliptical confidence regions for the transformed latent fields; these are then transformed from R^d to $(0,1)^D$ using the inverse link-function, giving confidence regions in compositional space. Having joint confidence regions for the compositions allowed us to evaluate the behaviour of all components as each individual component attains their lower and upper bounds in the confidence regions.

The statistical model was used to reconstruct past land-cover composition over Europe for five time periods using PbLCC data (Trondman et al., 2015) obtained from the REVEALS model (Sugita, 2007b). The land-cover reconstructions for the most recent time period were evaluated against present-time forest maps, and reconstructions for all time-periods were evaluated using cross-validation. The evaluations showed that a model containing both explanatory covariates and spatial dependence structure outperformed a model with only covariates, indicating that the addition of a spatial random effect improves predictions. Evaluations using the present-time forest maps showed that a model with

Dirichlet observations outperformed previously developed models using Gaussian observations of transformed fields (Pirzamanbein et al., 2014).

The reconstructed maps of land-cover composition can be used both in studies of climate models and to analyse changes in land-cover composition during the past millennia. For example, Fig. 6 uses the compositional distances (16) to illustrate the changes in land-cover composition between the five time periods considered in this study. This simple analysis shows that the largest changes in land cover between 4000 BCE and 1900 CE have occurred in Switzerland and Central France; along the North Sea coast in the UK, the Low Countries, Denmark, and southern Norway; and along the south Baltic coast in northern Germany and Poland.

The reconstructions of past land-cover composition obtained here are encouraging, as they clearly show the ability to recover continuous maps of past land cover from PbLCC data. The reconstructions from the full spatial model appear to conserve the information and trends from the pollen-based REVEALS estimates of past land cover (as discussed in Trondman et al., 2015) the best. They are also clearly better than previous spatial reconstructions in terms of e.g. the degree of openness and tree cover in the northernmost parts of Europe and the western coasts of Norway. Our future goal is to use these land-cover reconstructions in climate modelling studies and to gain insight into the effect of past anthropogenic deforestation. It is outside the scope of this paper to provide a discussion of the land-cover reconstructions in terms of historical changes in vegetation abundance, land cover and human impact over the past 6 000 years. However, the methods developed here provide (some of) the tools needed for such a discussion.

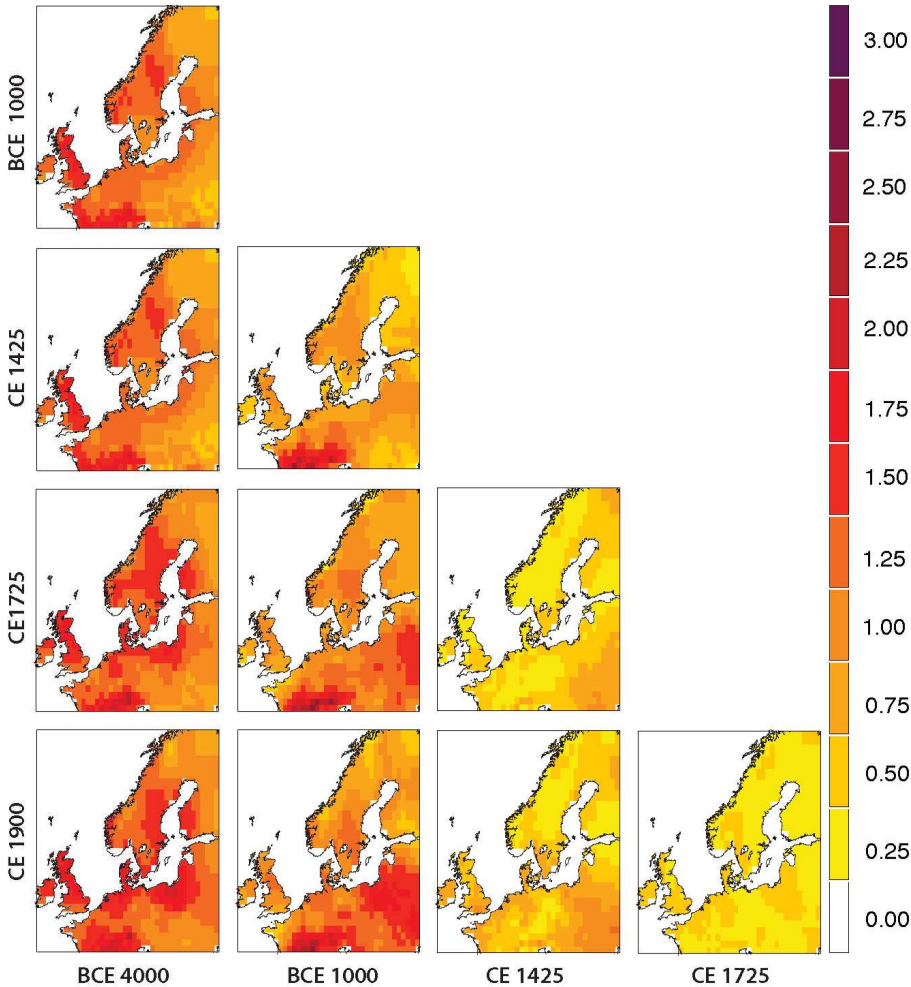


Figure 6: Compositional distances between the Full model land-cover reconstructions for the different time periods.

A Derivatives and Fisher Information of $[\boldsymbol{\eta}_{all}, \alpha|\mathbf{Y}]$

To construct the MALA updates for $\boldsymbol{\eta}_{all} = \mathbf{B}\boldsymbol{\beta} + \mathbf{X}, \alpha$ we need the derivatives and Fisher information of the log-posterior (8),

$$\begin{aligned} l(\mathbf{X}, \boldsymbol{\beta}, \alpha|\mathbf{Y}) &= \log\left(\prod_{s=1}^{N_o} \mathbb{P}(\mathbf{y}_s|f_s(\mathbf{A}\boldsymbol{\eta}_{all}), \alpha)\mathbb{P}(\mathbf{X}|\kappa, \boldsymbol{\rho})\mathbb{P}(\boldsymbol{\beta})\mathbb{P}(\alpha)\right) \\ &= \sum_{s=1}^{N_o} \log\Gamma(\alpha) - \sum_{s=1}^{N_o}\sum_{k=1}^{D} \log\Gamma(\alpha z_{s,k}) \\ &\quad + \sum_{s=1}^{N_o}\sum_{k=1}^{D} (\alpha z_{s,k} - 1)\log y_{s,k} - \frac{1}{2}\mathbf{X}^\top\left(\boldsymbol{\rho}^{-1} \otimes \mathbf{Q}(\kappa)\right)\mathbf{X} \\ &\quad - \frac{q_{\boldsymbol{\beta}}}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + (a_\alpha - 1)\log(\alpha) - \alpha b_\alpha + \text{const}. \end{aligned} \tag{17}$$

Here $\mathbf{A}\boldsymbol{\eta}_{all}$ is the latent $\mathbb{R}^d$-field at observed locations, $\{u_s\}_{s=1}^{N_o}$, $\mathbf{z}_s$ is the corresponding D-composition (i.e. defined on $(0,1)^D$, with $d = D-1$), and const is an additive constant. Before computing derivatives of the log-posterior, $l(\mathbf{X}, \boldsymbol{\beta}, \alpha|\mathbf{Y})$, we need some results for the compositional transformation.

A.1 Derivatives of Compositional Transforms

The compositional transform used in this paper is the additive log-ratio, (3), with inverse

$$z_k = \begin{cases} \frac{\exp(\eta_k)}{1+\sum_k^d \exp(\eta_k)}, & \text{if } k = 1, \ldots, D-1 \\ \frac{1}{1+\sum_k^d \exp(\eta_k)}, & \text{if } k = D. \end{cases} \tag{18}$$

Here z is a D-compositional value (i.e. $(0,1)^D$) and η is $\mathbb{R}^{D-1}$.

For the MALA-computations the first and second derivatives of the inverse transformation are needed. These can be expressed in terms of the compositions, z_k; for the first derivatives

$$\frac{\partial z_k}{\partial \eta_i} = \begin{cases} z_k(1 - z_k) & \text{if } k = i, \\ -z_k z_i & \text{if } k \neq i. \end{cases} \tag{19}$$

and for the second derivatives

$$\frac{\partial^2 z_k}{\partial \eta_i \partial \eta_j} = \begin{cases} z_k(1-z_k)(1-2z_k), & \text{if } i=j, k=i \\ -z_k z_i(1-2z_i), & \text{if } i=j, k\neq i \\ -z_j z_k(1-2z_k), & \text{if } i\neq j, k=i \\ 2z_k z_i z_j, & \text{if } i\neq j, k\neq i, k\neq j, \end{cases} \tag{20}$$

the case $i \neq j, k = j$ is obtained by symmetry.

One consequence of the sum to one constraint of compositional data is that the derivatives (19) and second derivatives (20) also sum to one:

$$\begin{aligned} \sum_{k=1}^{D} \frac{\partial z_k}{\partial \eta_i} &= z_i(1-z_i) - \sum_{k\neq i} z_i z_k = z_i\left(1 - \sum_{k=1}^{D} z_k\right) = 0 \\ \sum_{k=1}^{D} \frac{\partial^2 z_k}{\partial \eta_i^2} &= z_i(1-z_i)(1-2z_i) - \sum_{k\neq i} z_k z_i(1-2z_i) = 0 \\ \sum_{k=1}^{D} \frac{\partial^2 z_k}{\partial \eta_i \partial \eta_j} &= -z_j z_i(1-2z_i) - z_i z_j(1-2z_j) + \sum_{k\neq i,j} 2z_k z_i z_j = 0 \end{aligned} \tag{21}$$

A.2 Derivative of $l(\mathbf{X}, \beta, \alpha|\mathbf{Y})$

Recall that the latent field $\boldsymbol{\eta}$ is a linear combination of the mean zero spatial field(s) $\mathbf{X}$ and the regression coefficients β given as

$$\boldsymbol{\eta} = \mathbf{A}\left(\begin{bmatrix} \mathbb{I} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \beta \end{bmatrix} \right) = \begin{bmatrix} \mathbf{A} & \mathbf{AB} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \beta \end{bmatrix}.$$

Therefore, the updates of $\boldsymbol{\eta}$ are done by updating the underlying fields and regression coefficients. Thus we need the derivatives of $l(\mathbf{X}, \beta, \alpha|\mathbf{Y})$, i.e. ∇l, w.r.t

$\boldsymbol{\theta} = \begin{bmatrix}\mathbf{X}^\top & \beta^\top\end{bmatrix}^\top$ and α.

$$\nabla_{\boldsymbol{\theta}} l(\mathbf{X}, \beta, \alpha|\mathbf{Y}) = \begin{bmatrix}\mathbf{A} & \mathbf{AB}\end{bmatrix}^\top \nabla_{\boldsymbol{\eta}} \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}), \alpha) - \begin{bmatrix}\left(\rho^{-1} \otimes \mathbf{Q}(\kappa)\right)\mathbf{X} \\ q_\beta \beta\end{bmatrix} \tag{22a}$$

$$\begin{aligned}\frac{\partial l(\mathbf{X}, \beta, \alpha|\mathbf{y})}{\partial \alpha} = &\sum_{s=1}^{N_o} \psi(\alpha) - \sum_{s=1}^{N_o}\sum_{k=1}^{D} z_{s,k}\psi(\alpha z_{s,k}) \\ &+ \sum_{s=1}^{N_o}\sum_{k=1}^{D} z_{s,k} \log y_{s,k} + \frac{a_\alpha - 1}{\alpha} - b_\alpha\end{aligned} \tag{22b}$$

where $\nabla_{\boldsymbol{\theta}} l$ is the gradient w.r.t. $\boldsymbol{\theta}$ (a Nd-column vector) and $\psi(\cdot)$ is the digamma function. The elements of the gradient $\nabla_{\boldsymbol{\eta}} \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}), \alpha)$ (a $N_o d$-column vector) are

$$\frac{\partial \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}), \alpha)}{\partial \eta_{s,k}} = \sum_{l=1}^{D}\Big(-\alpha\psi(\alpha z_{s,l}) + \alpha \log y_{s,l}\Big)\frac{\partial z_{s,l}}{\partial \eta_{s,k}},$$

where the derivatives, $\partial z_{s,l}/\partial \eta_{s,k}$, depend on the choice of link function (see (19) for the additive log ratio case).

A.3 The Fisher Information

The Fisher information used in the MALA updates is computed as the expectation of the Hessian over observations, $\mathbf{Y}$, given all parameters and latent fields:

$$\mathcal{I} = -\mathsf{E}_{\mathbf{Y}}\big(\mathbf{H}(l) \,|\, \mathbf{X}, \beta, \alpha, \rho, \kappa\big) = \begin{bmatrix}\mathcal{I}_{\boldsymbol{\theta},\boldsymbol{\theta}} & \mathcal{I}_{\boldsymbol{\theta},\alpha} \\ \mathcal{I}_{\alpha,\boldsymbol{\theta}} & \mathcal{I}_{\alpha,\alpha}\end{bmatrix} \tag{23}$$

where $\mathbf{H}(l)$ is Hessian of $l(\mathbf{X}, \beta, \alpha|\mathbf{Y})$. The resulting matrix consists of four blocks: two with second derivatives w.r.t. $\boldsymbol{\theta}$ and α, and two with cross partial derivatives; each of the blocks is described below. For brevity we use $\mathsf{E}(\mathbf{H}(l) \,|\, \bullet)$ to denote the conditional expectation in (23), and note that

$$\mathsf{E}\big(\log y_{s,k} \,\big|\, \bullet\big) = \psi\left(\alpha z_{s,k}\right) - \psi\left(\sum_{l=1}^{D} \alpha z_{s,l}\right) = \psi(\alpha z_{s,k}) - \psi(\alpha). \tag{24}$$

Similar to (22a) the top left block can be written as

$$\mathcal{I}_{\boldsymbol{\theta},\boldsymbol{\theta}} = \begin{bmatrix}\mathbf{A} & \mathbf{AB}\end{bmatrix}^\top \mathbf{H}_{\boldsymbol{\eta}} \begin{bmatrix}\mathbf{A} & \mathbf{AB}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\rho}^{-1} \otimes \mathbf{Q}(\kappa) & \mathbf{0} \\ \mathbf{0} & q_{\boldsymbol{\beta}}\mathbb{I}\end{bmatrix},$$

where $\mathbf{H}_{\boldsymbol{\eta}}$ is a symmetric $N_o d \times N_o d$ matrix with elements

$$\mathbf{H}_{\boldsymbol{\eta}}^{(sk,s'k')} = -\mathsf{E}\left(\frac{\partial^2 \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}),\alpha)}{\partial \eta_{s,k}\partial \eta_{s',k'}} \,\middle|\, \bullet\right).$$

The elements in $\mathbf{H}_{\boldsymbol{\eta}}$ are indexed by their spatial location, $s = 1,\ldots,N_o$, and which latent field, $k = 1,\ldots,d$, they belong to (i.e. which transformed compositional component). For elements at different locations

$$\mathbf{H}_{\boldsymbol{\eta}}^{(sk,s'k')} = 0, \quad \text{if} \quad s \neq s',$$

leaving only

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\eta}}^{(sk,sk')} =& -\mathsf{E}\left(\frac{\partial^2 \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}),\alpha)}{\partial \eta_{s,k}\partial \eta_{s,k'}} \,\middle|\, \bullet\right) \\
=& -\frac{\partial}{\partial \eta_{s,k'}}\mathsf{E}\left(\sum_{l=1}^{D}\left(-\alpha\psi(\alpha z_{s,l}) + \alpha \log y_{s,l}\right)\frac{\partial z_{s,l}}{\partial \eta_{s,k}} \,\middle|\, \bullet\right) \\
=& \alpha^2 \sum_{l=1}^{D} \psi'(\alpha z_{s,l}) \frac{\partial z_{s,l}}{\partial \eta_{s,k'}}\frac{\partial z_{s,l}}{\partial \eta_{s,k}} \\
&+ \alpha \sum_{l=1}^{D}\left(\psi(\alpha z_{s,l}) - \mathsf{E}\left(\log y_{s,l} \,\middle|\, \bullet\right)\right)\frac{\partial^2 z_{s,l}}{\partial \eta_{s,k}\partial \eta_{s,k'}}.
\end{aligned}$$

Using the expectations in (24) gives

$$\mathbf{H}_{\boldsymbol{\eta}}^{(sk,sk')} = \alpha^2 \sum_{l=1}^{D} \psi'(\alpha z_{s,l}) \frac{\partial z_{s,l}}{\partial \eta_{s,k'}}\frac{\partial z_{s,l}}{\partial \eta_{s,k}} + \alpha\psi(\alpha) \sum_{l=1}^{D} \frac{\partial^2 z_{s,l}}{\partial \eta_{s,k}\partial \eta_{s,k'}},$$

and with the sum to zero result in (21) the elements of $\mathbf{H}_{\boldsymbol{\eta}}$ simplify to

$$\mathbf{H}_{\boldsymbol{\eta}}^{(sk,sk')} = \alpha^2 \sum_{l=1}^{D} \psi'(\alpha z_{s,l}) \frac{\partial z_{s,l}}{\partial \eta_{s,k'}}\frac{\partial z_{s,l}}{\partial \eta_{s,k}}.$$

Derivation of (22a) w.r.t. α gives

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta},\alpha} = -\begin{bmatrix}\mathbf{A} & \mathbf{AB}\end{bmatrix}^{\top} \mathsf{E}\left(\frac{\partial}{\partial\alpha}\nabla_{\boldsymbol{\eta}} \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}),\alpha) \,\middle|\, \bullet\right),$$

since only the Dirichlet part of the log-likelihood contributes too the cross-derivatives. The part concerning the gradient, $\nabla_{\boldsymbol{\eta}} \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}),\alpha)$, gives a column vector of length $N_o d$ with elements

$$\begin{aligned}\mathsf{E}\left(\frac{\partial^2 \log \mathbb{P}(\mathbf{Y}|f(\boldsymbol{\eta}),\alpha)}{\partial\alpha\partial\eta_{s,k}} \,\middle|\, \bullet\right) &= \sum_{l=1}^{D}\Big(-\psi(\alpha z_{s,l}) - \alpha z_{s,l}\psi'(\alpha z_{s,l}) + \\ &\qquad + \mathsf{E}\big(\log y_{s,l} \,\big|\, \bullet\big)\Big)\frac{\partial z_{s,l}}{\partial\eta_{s,k}}, \\ &= -\alpha\sum_{l=1}^{D} z_{s,l}\psi'(\alpha z_{s,l})\frac{\partial z_{s,l}}{\partial\eta_{s,k}}.\end{aligned}$$

The last equality is obtained from (24) and (21). Symmetry gives that $\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta},\alpha} = \boldsymbol{\mathcal{I}}_{\alpha,\boldsymbol{\theta}}^{\top}$.

The last block of (23) is

$$\begin{aligned}\mathcal{I}_{\alpha,\alpha} &= -\mathsf{E}\left(\frac{\partial^2 l(\mathbf{X},\beta,\alpha|\mathbf{y})}{\partial\alpha^2} \,\middle|\, \bullet\right) \\ &= -\left(\sum_{s=1}^{N_o}\psi(\alpha) - \sum_{s=1}^{N_o}\sum_{k=1}^{D} z_{s,k}^2\psi(\alpha z_{s,k}) - \frac{a_\alpha - 1}{\alpha^2}\right).\end{aligned}$$

B The Posterior $\kappa|\mathbf{X}$

The posterior of $\kappa|\mathbf{X}$ is obtained by integrating out ρ from the joint posterior of $\kappa, \rho|\mathbf{X}$. With the densities for $\mathbf{X}$ and ρ given as

$$\mathbf{X}|\kappa, \rho \sim \mathsf{N}\left(\mathbf{0}, \rho \otimes \mathbf{Q}^{-1}(\kappa)\right) \qquad \text{and} \qquad \rho \sim \mathsf{IW}\left(a_\rho \mathbb{I}, b_\rho\right) \tag{25}$$

in (7) the posterior $\kappa|\mathbf{X}$ is

$$\begin{aligned}\mathbb{P}(\kappa|\mathbf{X}) &\propto \int \mathbb{P}(\mathbf{X}|\kappa, \rho)\mathbb{P}(\kappa)\mathbb{P}(\rho)d\rho \\ &\propto \int \left|\rho^{-1} \otimes \mathbf{Q}(\kappa)\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{X}^\top \left(\rho^{-1} \otimes \mathbf{Q}(\kappa)\right)\mathbf{X}\right)\mathbb{P}(\kappa) \\ &\quad \cdot \left|a_\rho \mathbb{I}\right|^{\frac{b_\rho}{2}} |\rho|^{-\frac{b_\rho+d+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\rho^{-1} a_\rho \mathbb{I}\right)\right) \mathrm{d}\rho.\end{aligned} \tag{26}$$

Introducing vectorization such that $\operatorname{vec}(\mathbf{x}) = \mathbf{X}$, where $\mathbf{x} = (X_1, \cdots, X_d)$ is a $N \times d$-matrix version of the column-vector $\mathbf{X}$, the exponential term can be rewritten as

$$\begin{aligned}&-\frac{1}{2}\mathbf{X}^\top \left(\rho^{-1} \otimes \mathbf{Q}(\kappa)\right)\mathbf{X} = -\frac{1}{2}\mathbf{X}^\top \operatorname{vec}\left(\mathbf{Q}(\kappa)^\top \mathbf{x} \rho^{-1}\right) \\ =&-\frac{1}{2}\operatorname{tr}\left(\mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}\rho^{-1}\right) = -\frac{1}{2}\operatorname{tr}\left(\rho^{-1}\mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}\right).\end{aligned}$$

The posterior in (26) now simplifies to

$$\begin{aligned}\mathbb{P}(\kappa|\mathbf{X}) \propto &\mathbb{P}(\kappa)\left|a_\rho \mathbb{I}\right|^{\frac{b_\rho}{2}} \left|\mathbf{Q}(\kappa)\right|^{\frac{d}{2}} \\ &\int |\rho|^{-\frac{N+b_\rho+d+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\rho^{-1}\left(a_\rho \mathbb{I} + \mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}\right)\right)\right) \mathrm{d}\rho.\end{aligned}$$

Recognizing the density of an unnormalized inverse-Wishart distribution under the integral sign we normalise and obtain the posteriors

$$\begin{aligned}&\rho|\kappa, \mathbf{X} \sim \mathsf{IW}\left(a_\rho \mathbb{I} + \mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}, b_\rho + N\right), \\ &\mathbb{P}(\kappa|\mathbf{X}) \propto \mathbb{P}(\kappa) \cdot \frac{a_\rho^{\frac{db_\rho}{2}} \left|\mathbf{Q}(\kappa)\right|^{\frac{d}{2}}}{\left|a_\rho \mathbb{I} + \mathbf{x}^\top \mathbf{Q}(\kappa)\mathbf{x}\right|^{\frac{N+b_\rho}{2}}}.\end{aligned}$$

C Parameter Estimates

Table 3: Parameter estimates (Est) and 95% quantile (CI) for the two models (Full — spatial model and RM — regression model) used to reconstruct past land-cover composition from the PbLCC data.

1700 CE

	Full		RM	
Parameter	Est	(CI)	Est	(CI)
α	9.55	(7.64 , 12.93)	6.07	(5.31 , 6.86)
κ	0.23	(0.12 , 0.39)	-	-
ρ_{11}	0.50	(0.13 , 1.83)	-	-
ρ_{12}	0.23	(0.01 , 1.11)	-	-
ρ_{22}	0.25	(0.07 , 0.90)	-	-
β_{10}	-0.72	(-1.85 , 0.23)	-0.13	(-0.24 , 0.00)
β_{11}	0.17	(0.08 , 0.26)	0.28	(0.25 , 0.31)
β_{12}	-0.01	(-0.12 , 0.10)	-0.08	(-0.12 , -0.03)
β_{13}	-0.01	(-0.20 , 0.18)	-0.14	(-0.24 , -0.05)
β_{20}	-0.74	(-1.56 , 0.02)	-0.35	(-0.49 , -0.23)
β_{21}	0.07	(-0.01 , 0.15)	0.13	(0.11 , 0.16)
β_{22}	-0.05	(-0.13 , 0.03)	-0.08	(-0.12 , -0.04)
β_{23}	-0.20	(-0.38 , -0.03)	-0.30	(-0.40 , -0.20)

Table 4: Parameter estimates (Est) and 95% quantile (CI) for the two models (Full — spatial model and RM — regression model) used to reconstruct past land-cover composition from the PbLCC data.

1400 CE

	Full		RM	
Parameter	Est	(CI)	Est	(CI)
α	8.75	(7.22 , 10.76)	5.18	(4.57 , 5.83)
κ	0.18	(0.08 , 0.31)	-	-
ρ_{11}	0.37	(0.10 , 0.98)	-	-
ρ_{12}	0.12	(-0.02 , 0.47)	-	-
ρ_{22}	0.17	(0.06 , 0.44)	-	-
β_{10}	-0.70	(-2.34 , 0.77)	-0.09	(-0.21 , 0.03)
β_{11}	0.16	(0.05 , 0.26)	0.28	(0.26 , 0.31)
β_{12}	0.02	(-0.10 , 0.13)	-0.07	(-0.12 , -0.02)
β_{13}	0.09	(-0.10 , 0.28)	-0.09	(-0.19 , 0.00)
β_{20}	-0.56	(-1.78 , 0.51)	-0.15	(-0.28 , -0.03)
β_{21}	0.07	(-0.03 , 0.16)	0.14	(0.11 , 0.17)
β_{22}	-0.03	(-0.12 , 0.06)	-0.07	(-0.11 , -0.03)
β_{23}	-0.18	(-0.36 , -0.01)	-0.32	(-0.42 , -0.22)

Table 5: Parameter estimates (Est) and 95% quantile (CI) for the two models (Full — spatial model and RM — regression model) used to reconstruct past land-cover composition from the PbLCC data.

1000 BCE

	Full		RM	
Parameter	Est	(CI)	Est	(CI)
α	7.02	(5.89 , 8.37)	4.42	(3.91 , 4.96)
κ	0.19	(0.08 , 0.30)	-	-
ρ_{11}	0.32	(0.10 , 0.80)	-	-
ρ_{12}	0.07	(-0.04 , 0.27)	-	-
ρ_{22}	0.16	(0.06 , 0.37)	-	-
β_{10}	0.19	(-1.40 , 1.57)	0.50	(0.37 , 0.63)
β_{11}	0.24	(0.13 , 0.35)	0.30	(0.27 , 0.33)
β_{12}	-0.03	(-0.17 , 0.10)	0.05	(-0.01 , 0.11)
β_{13}	0.15	(-0.06 , 0.37)	-0.02	(-0.12 , 0.09)
β_{20}	0.19	(-1.07 , 1.21)	0.55	(0.43 , 0.68)
β_{21}	0.07	(-0.01 , 0.16)	0.12	(0.09 , 0.14)
β_{22}	0.03	(-0.08 , 0.13)	0.02	(-0.03 , 0.07)
β_{23}	-0.02	(-0.20 , 0.18)	-0.11	(-0.22 , -0.01)

Table 6: Parameter estimates (Est) and 95% quantile (CI) for the two models (Full — spatial model and RM — regression model) used to reconstruct past land-cover composition from the PbLCC data.

4000 BCE

	Full		RM	
Parameter	Est	(CI)	Est	(CI)
α	7.58	(6.26 , 9.84)	5.36	(4.72 , 6.02)
κ	0.20	(0.10 , 0.32)	-	-
ρ_{11}	0.21	(0.07 , 0.69)	-	-
ρ_{12}	0.10	(-0.02 , 0.64)	-	-
ρ_{22}	0.24	(0.06 , 1.03)	-	-
β_{10}	0.41	(-0.70 , 1.48)	0.38	(0.24 , 0.53)
β_{11}	0.23	(0.13 , 0.33)	0.23	(0.20 , 0.26)
β_{12}	-0.19	(-0.36 , -0.03)	-0.04	(-0.11 , 0.03)
β_{13}	0.04	(-0.17 , 0.25)	-0.04	(-0.14 , 0.07)
β_{20}	0.61	(-0.66 , 1.58)	0.99	(0.85 , 1.12)
β_{21}	0.01	(-0.09 , 0.10)	0.08	(0.06 , 0.11)
β_{22}	-0.01	(-0.16 , 0.14)	0.00	(-0.06 , 0.05)
β_{23}	-0.05	(-0.24 , 0.15)	-0.25	(-0.35 , -0.15)

D Maps of Estimated Land Cover

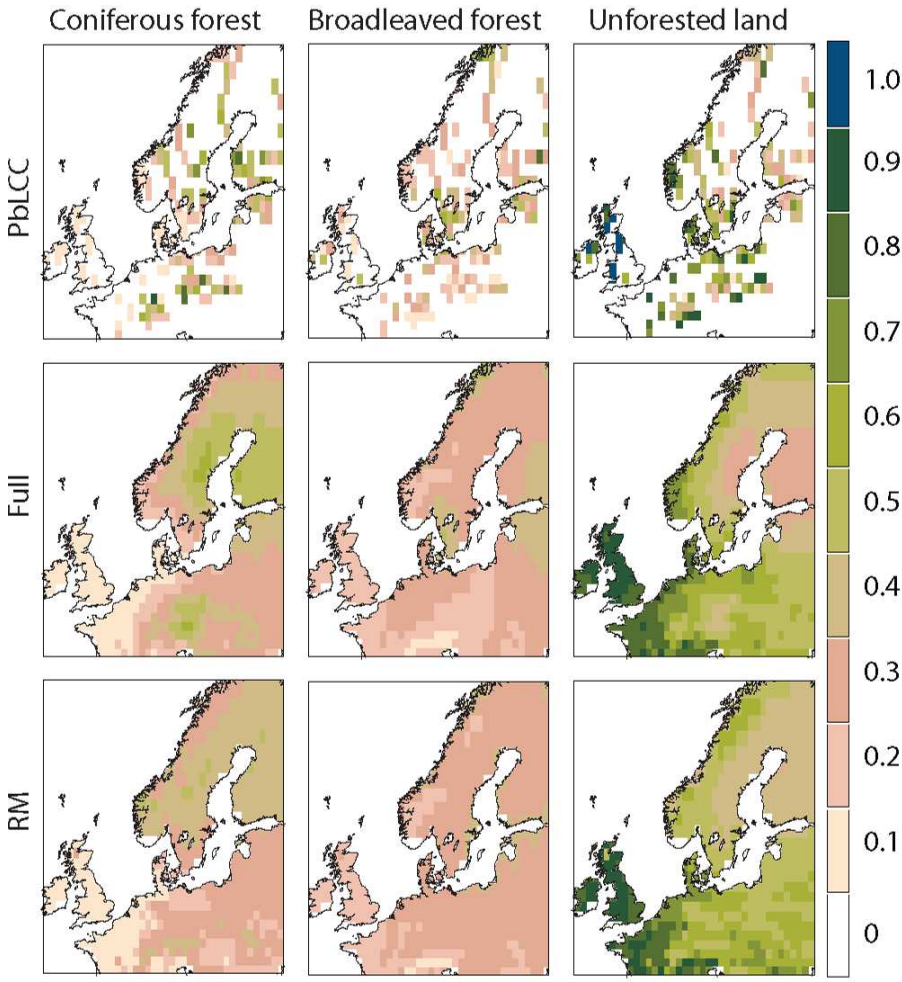


Figure 7: Results for the 1725 CE time period: the first row shows the PbLCC data, and the other rows show the reconstructions for the full spatial model (Full) and the regression model (RM).

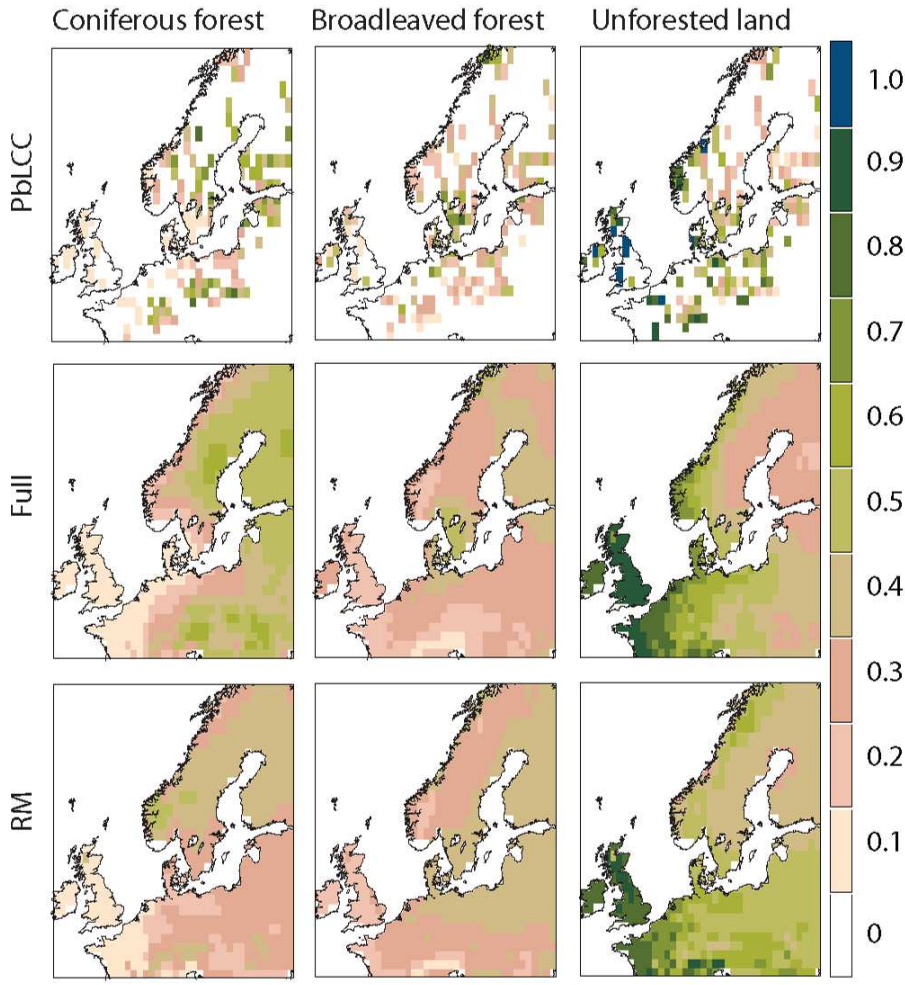


Figure 8: Results for the 1425 CE time period: the first row shows the PbLCC data, and the other rows show the reconstructions for the full spatial model (Full) and the regression model (RM).

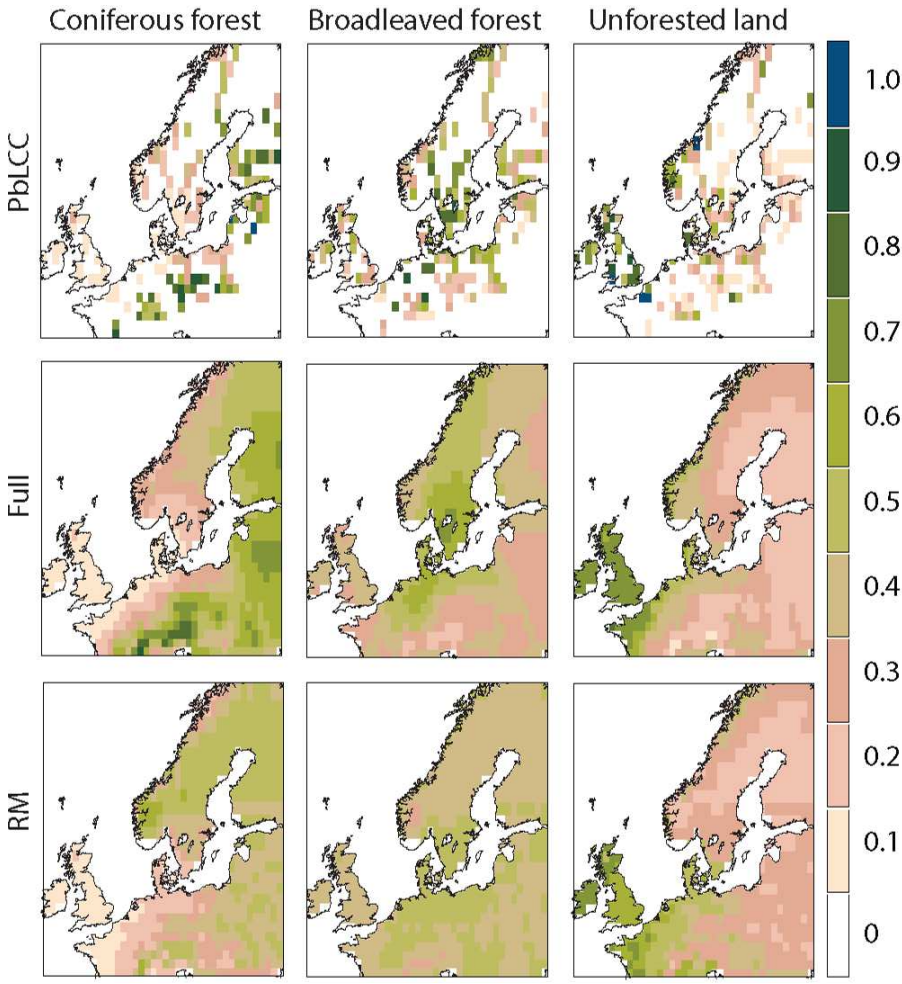


Figure 9: Results for the 1000 BCE time period: the first row shows the PbLCC data, and the other rows show the reconstructions for the full spatial model (Full) and the regression model (RM).

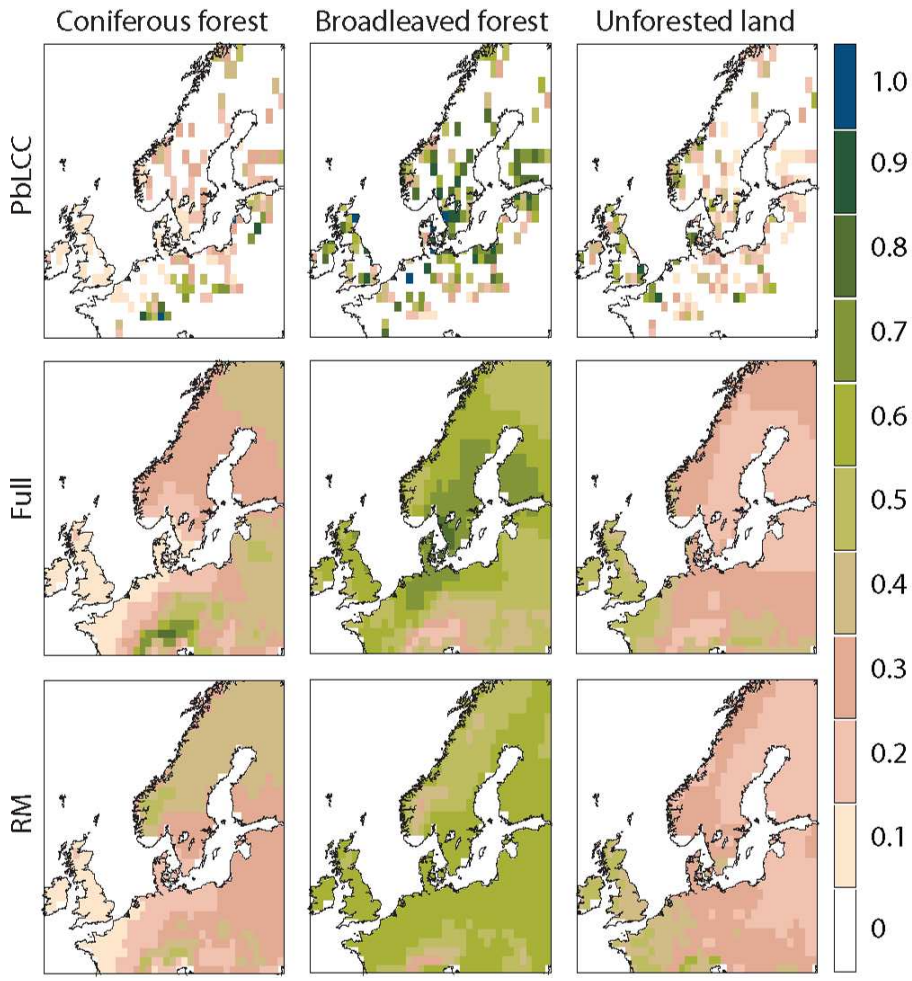


Figure 10: Results for the 4000 BCE time period: the first row shows the PbLCC data, and the other rows show the reconstructions for the full spatial model (Full) and the regression model (RM).

E Uncertainties in Estimated Land Cover

E.1 Maps of Uncertainties

This appendix contains figures illustrating the predication uncertainties for all five time periods. All figures contain:
The first column shows the reconstructed land-cover composition for the time period, using the full spatial model. Columns 2 and 3, row 1 (with thick/red axes), show the maximum and minimum of 95% elliptical confidence regions for Coniferous; rows 2 and 3 give the corresponding Broadleaved and Unforested compositions. Columns 4 and 5 (row 2 with thick/red axes) gives the bounds for the Broadleaved composition while columns 6 and 7 show the bounds for Unforested land (row 3 with thick/red axes). The concept of joint confidence interval for compositions is illustrated in Fig 2.

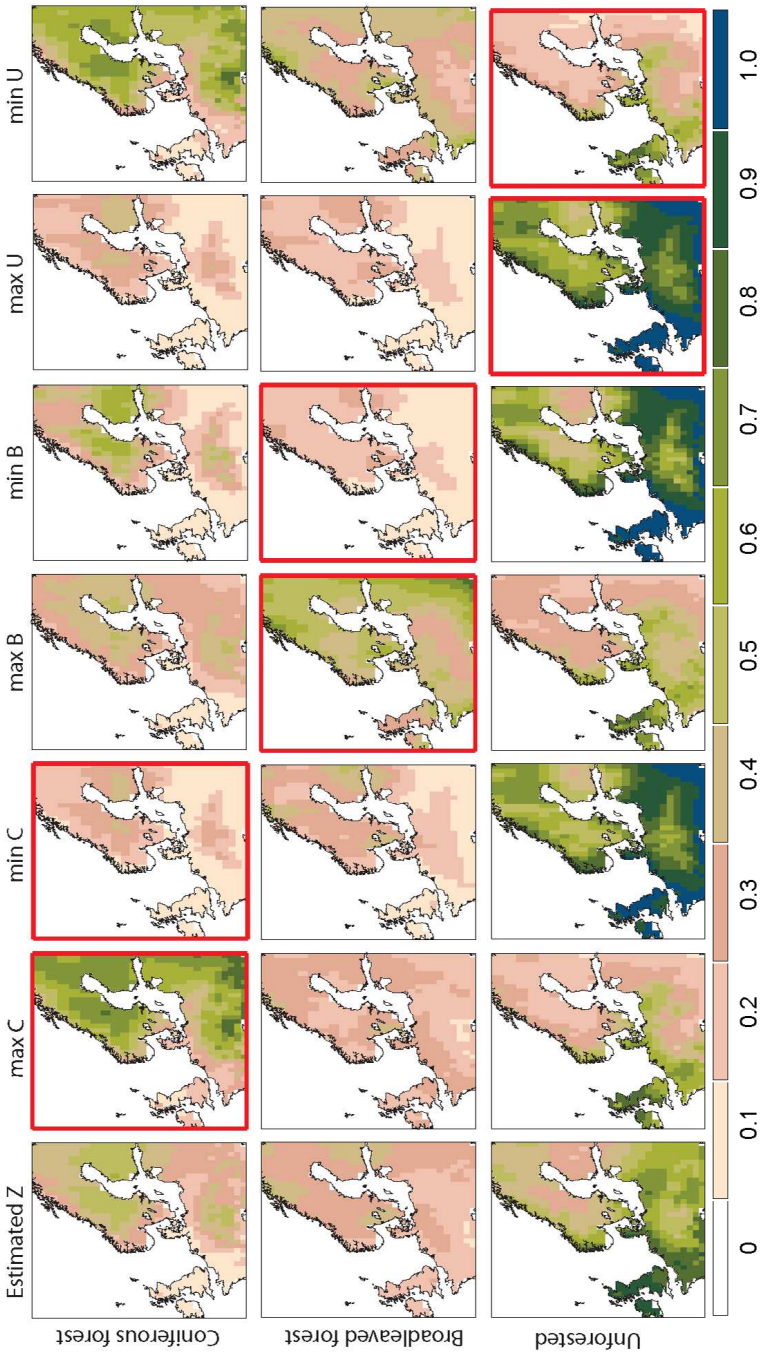


Figure 11: Reconstructions for 1900 CE. For a description of the figure see 112.

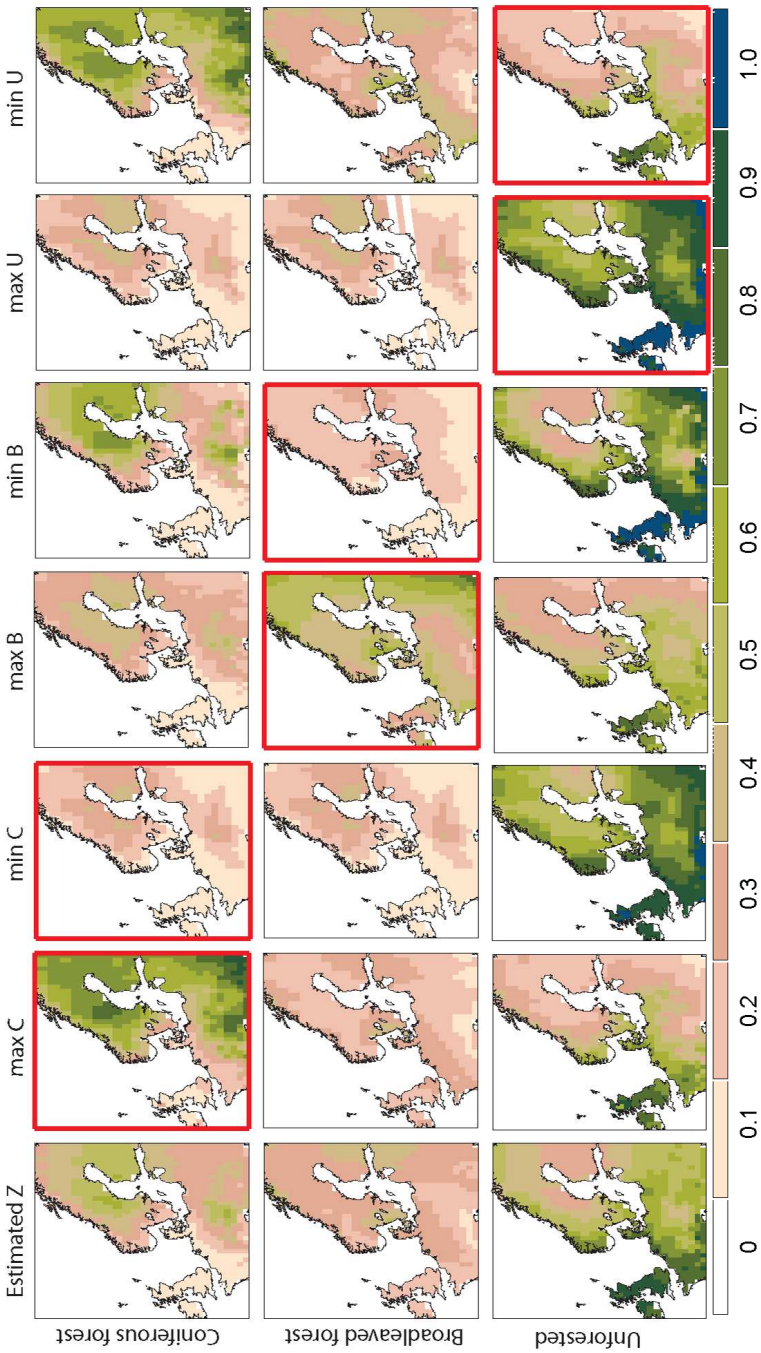


Figure 12: Reconstructions for 1725 CE. For a description of the figure see 112.

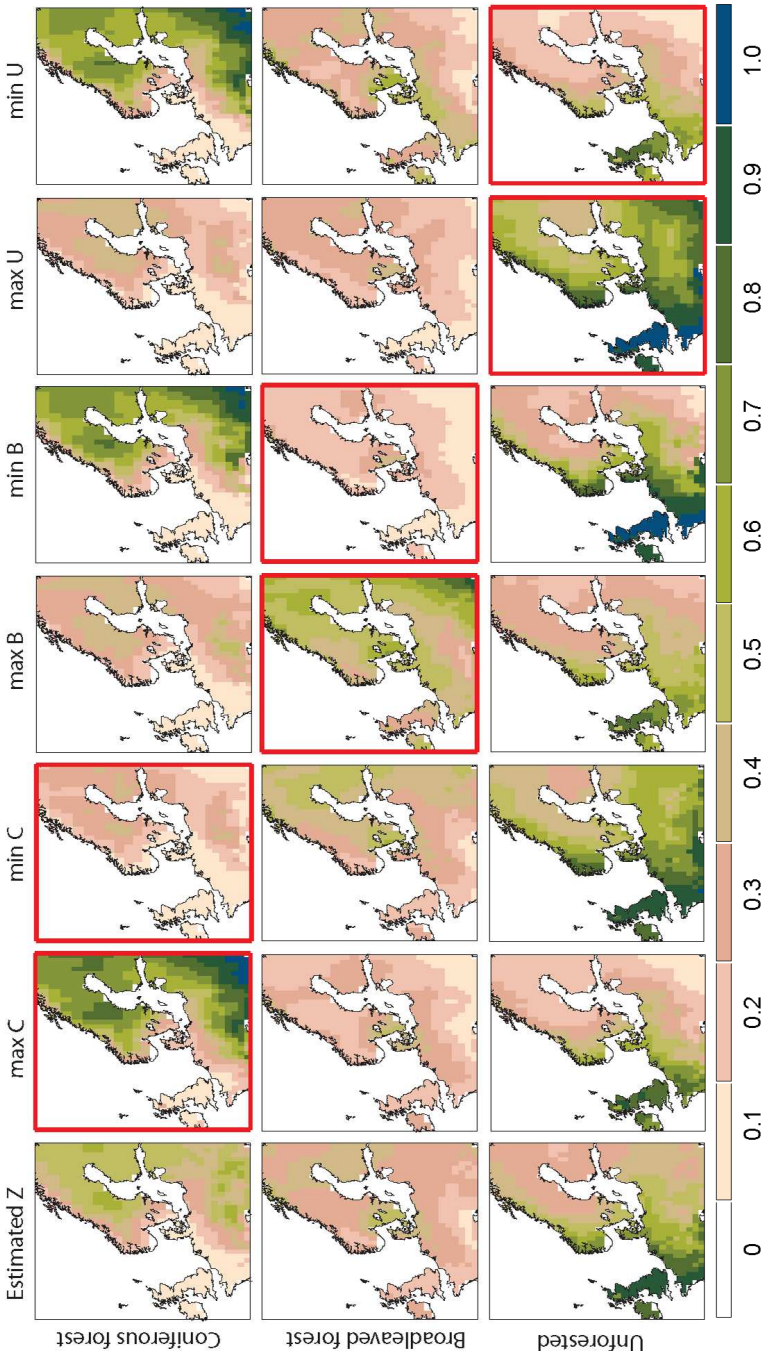


Figure 13: Reconstructions for 1425 CE. For a description of the figure see 112.

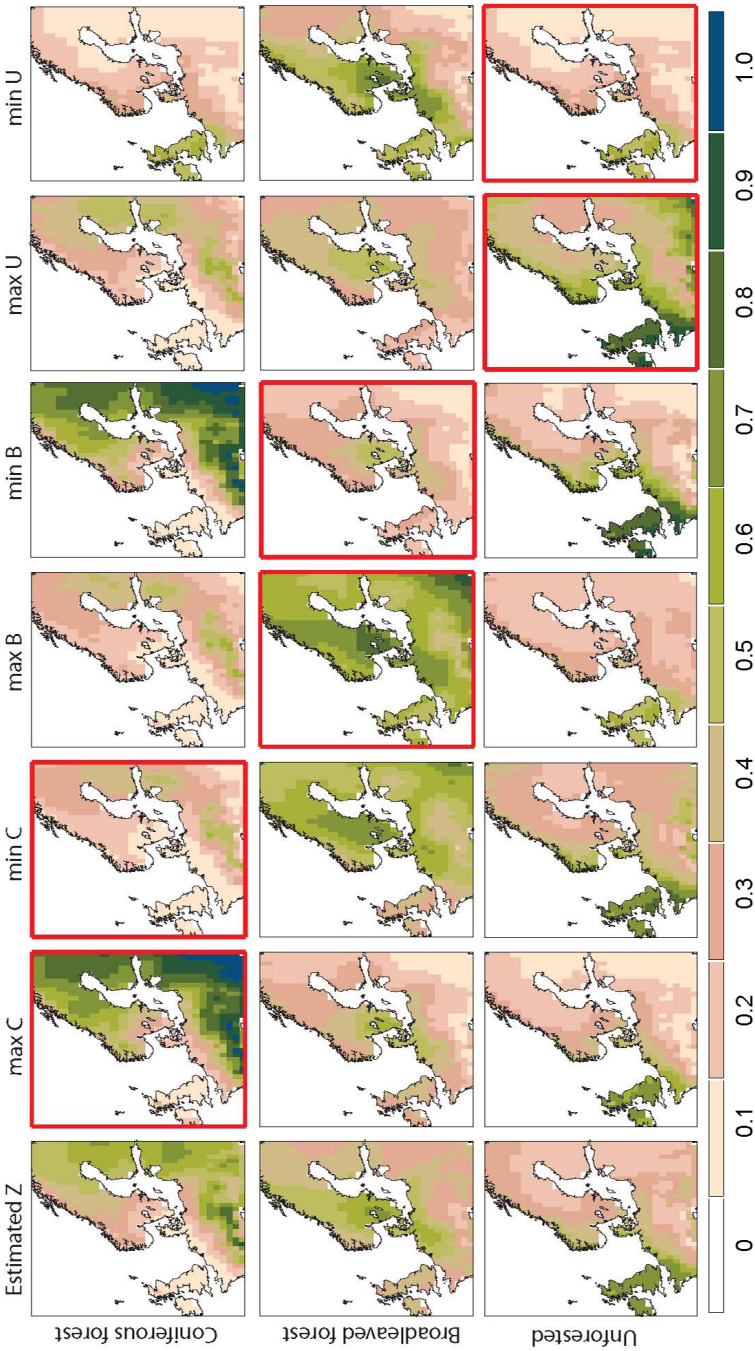


Figure 14: Reconstructions for 1000 BCE. For a description of the figure see 112.

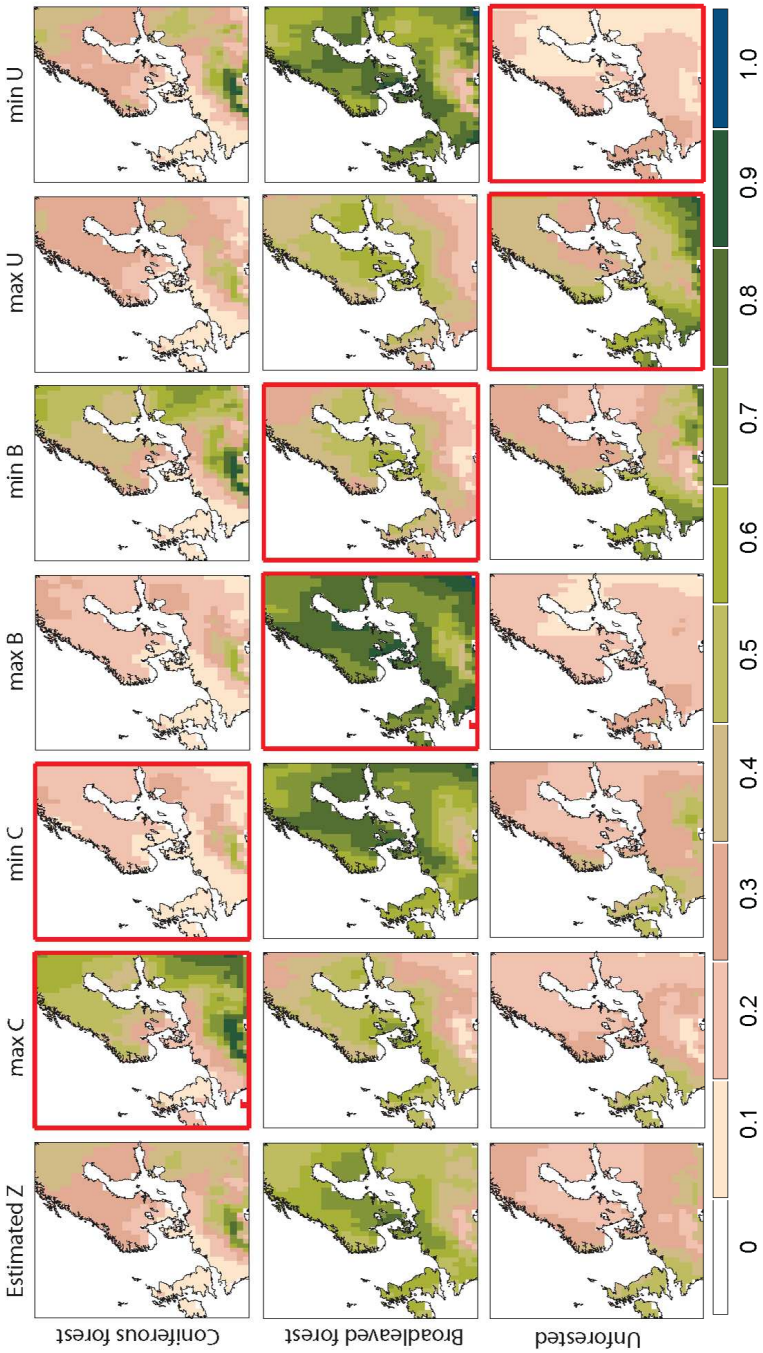


Figure 15: Reconstructions for 4000 BCE. For a description of the figure see 112.

E.2 Confidence Regions for Selected Locations

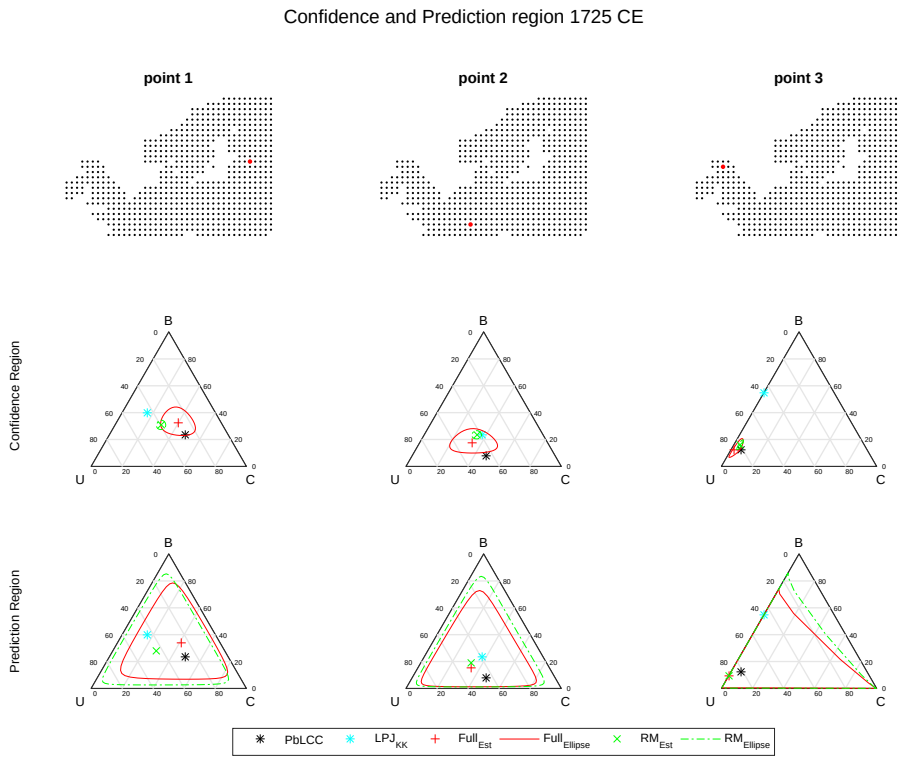


Figure 16: The first row shows the locations of the three selected grid cells for the 1725 CE time period. The second row shows the ternary confidence regions along with the reconstructions for the two models (Full—spatial model; RM—regression model) and the values of the PbLCC data and the LPJ-GUESS$_{KK}$ land cover covariate at each location. The third row shows the ternary prediction regions and the same values.

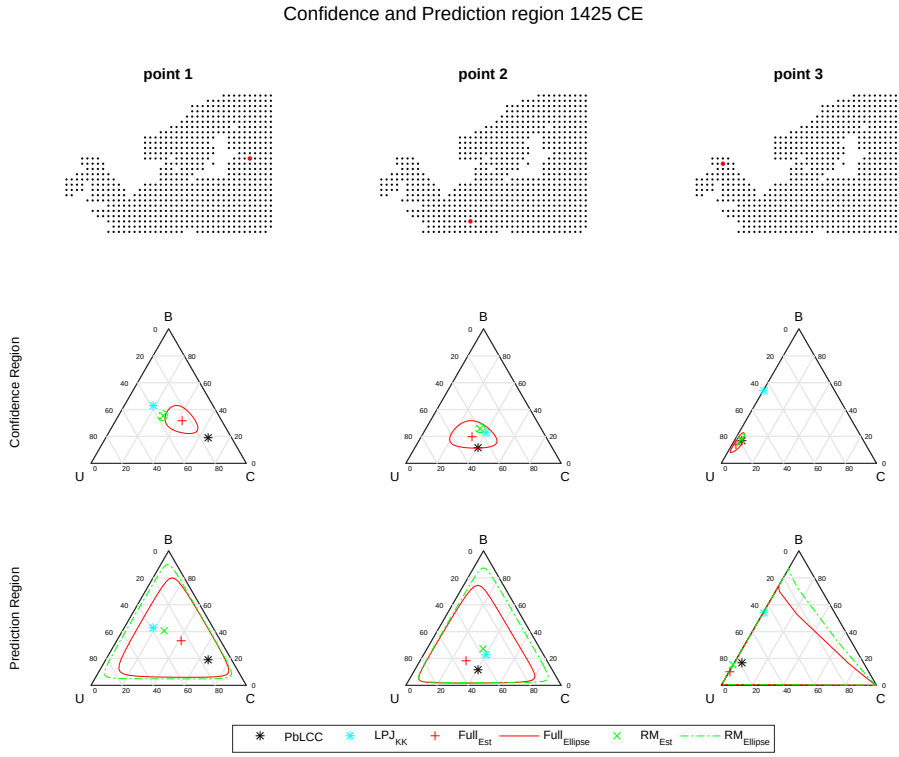


Figure 17: The first row shows the locations of the three selected grid cells for the 1425 CE time period. The second row shows the ternary confidence regions along with the reconstructions for the two models (Full—spatial model; RM—regression model) and the values of the PbLCC data and the LPJ-GUESS$_{KK}$ land cover covariate at each location. The third row shows the ternary prediction regions and the same values.

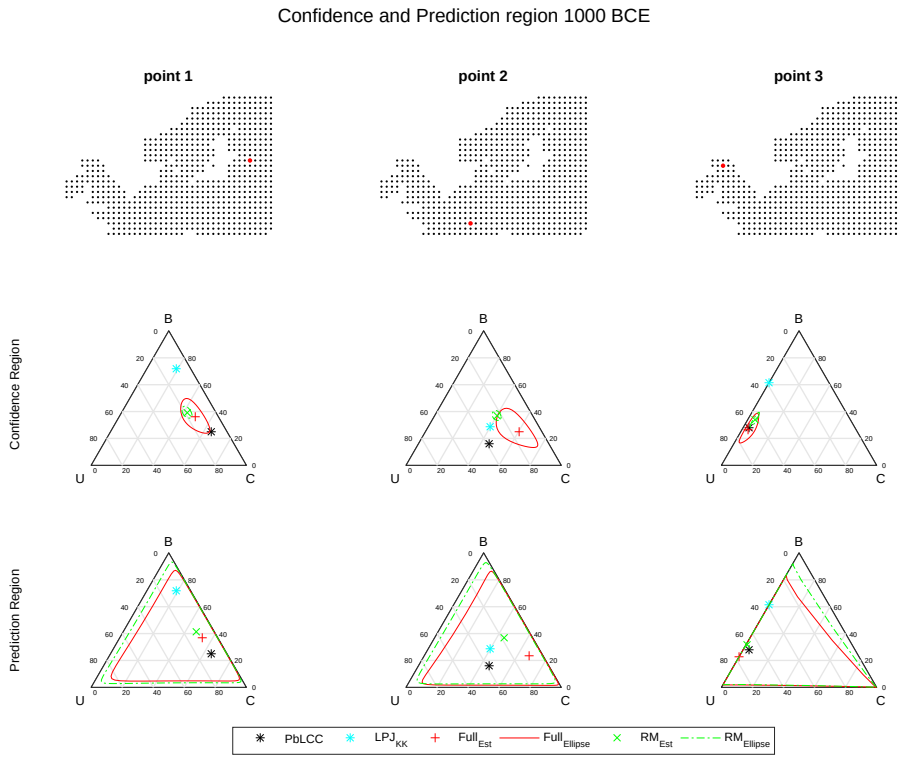


Figure 18: The first row shows the locations of the three selected grid cells for the 1000 BCE time period. The second row shows the ternary confidence regions along with the reconstructions for the two models (Full—spatial model; RM—regression model) and the values of the PbLCC data and the LPJ-GUESS$_{KK}$ land cover covariate at each location. The third row shows the ternary prediction regions and the same values.

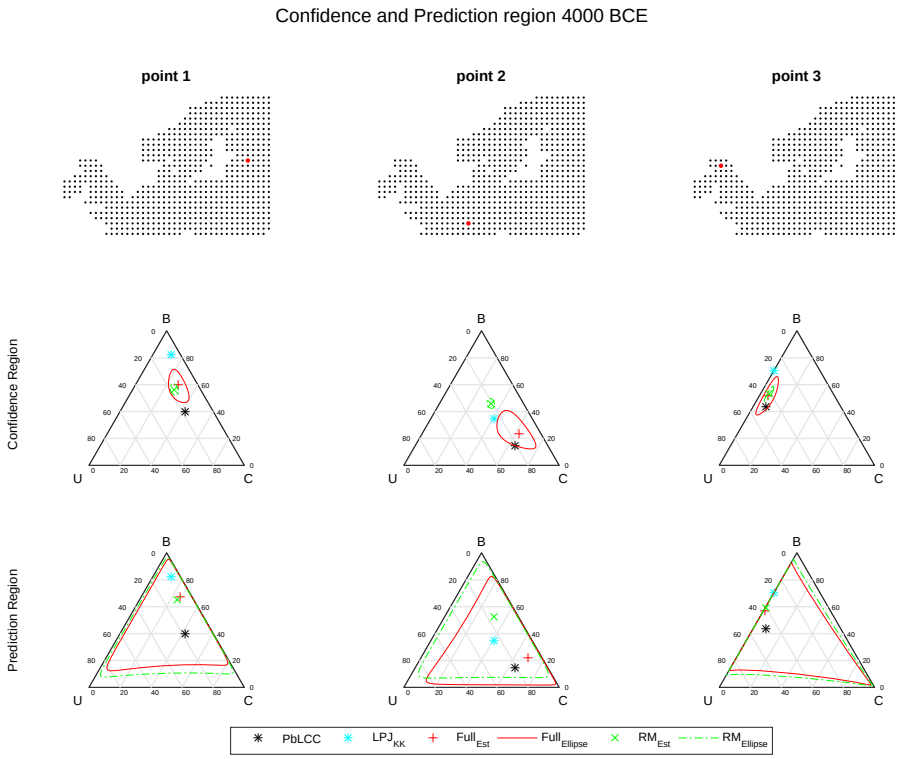


Figure 19: The first row shows the locations of the three selected grid cells for the 4000 BCE time period. The second row shows the ternary confidence regions along with the reconstructions for the two models (Full—spatial model; RM—regression model) and the values of the PbLCC data and the LPJ-GUESS$_{KK}$ land cover covariate at each location. The third row shows the ternary prediction regions and the same values.

References

J. Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., 1986.

J. Aitchison. On criteria for measures of compositional difference. *Math. Geol.*, 24(4):365–379, 1992.

J. Aitchison, C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn. Logratio analysis and compositional distance. *Math. Geol.*, 32(3):271–275, 2000.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. and Comput.*, 18(4):343–373, 2008.

J. J. Becker, D. T. Sandwell, W. H. F. Smith, J. Braud, B. Binder, J. Depner, D. Fabre, J. Factor, S. Ingalls, S. H. Kim, R. Ladner, K. Marks, S. Nelson, A. Pharaoh, G. Sharman, R. Trimmer, J. VonRosenburg, G. Wallace, and P. Weatherall. Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Mar. Geod.*, 32(4):355–371, 2009.

D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *J. Am. Statist. Assoc.*, 96(456):1205–1214, 2001.

V. Brovkin, M. Claussen, E. Driesschaert, T. Fichefet, D. Kicklighter, M. Loutre, H. Matthews, N. Ramankutty, M. Schaeffer, and A. Sokolov. Biogeophysical effects of historical land cover changes simulated by six earth system models of intermediate complexity. *Clim. Dynam.*, 26(6):587–600, 2006.

M. Claussen, V. Brovkin, and A. Ganopolski. Biogeophysical versus biogeochemical feedbacks of large-scale land cover change. *Geophys. Res. Lett.*, 28(6):1011–1014, 2001.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Math. Geol.*, 35(3):279–300, 2003.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Interpretable priors for hyperparameters for Gaussian Random Fields. Technical Report 1503.00256v2, arXiv, 2016. URL `http://arxiv.org/abs/1503.00256v2`.

M.-J. Gaillard, S. Sugita, F. Mazier, A.-K. Trondman, A. Brostrom, T. Hickler, J. O. Kaplan, E. Kjellström, U. Kokfelt, P. Kuneš, , C. Lemmen, P. Miller, J. Olofsson, A. Poska, M. Rundgren, B. Smith, G. Strandberg, R. Fyfe, A. Nielsen, T. Alenius, L. Balakauskas, L. Barnekov, H. Birks, A. Bjune, L. Björkman, T. Giesecke, K. Hjelle, L. Kalnina, M. Kangur, W. van der Knaap, T. Koff, P. Lagerås, M. Latałowa, M. Leydet, J. Lechterbeck, M. Lindbladh, B. Odgaard, S. Peglar, U. Segerström, H. von Stedingk, and H. Seppä. Holocene land-cover reconstructions for studies on land cover-climate feedbacks. *Clim. Past.*, 6:483–499, 2010.

M.-J. Gaillard, T. Kleinen, P. Samuelsson, A.-B. Nielsen, J. Bergh, J. Kaplan, A. Poska, C. Sandström, G. Strandberg, A.-K. Trondman, and A. Wramneby. Causes of regional change-land cover. In *Second Assessment of Climate Change for the Baltic Sea Basin*, Regional Climate Studies, pages 453–477. Springer International Publishing, 2015.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 73(2):123–214, 2011.

G. H. Givens and J. A. Hoeting. *Computational statistics*, volume 710. John Wiley & Sons, 2012.

M. Haran. Gaussian Random Field Models for spatial data. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 449–478. Chapman & Hall/CRC, 2011.

S. Hellman, M.-j. Gaillard, A. Broström, and S. Sugita. Effects of the sampling design and selection of parameter values on pollen-based quantitative reconstructions of regional vegetation: a case study in southern Sweden using the REVEALS model. *Veg. Hist. Archaeobot.*, 17(5):445–459, 2008.

J. O. Kaplan, K. M. Krumhardt, and N. Zimmermann. The prehistoric and preindustrial deforestation of Europe. *Quaternary. Sci. Rev.*, 28(27):3016–3034, 2009.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Stat.*, 41(2):495–502, 1970.

K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Global. Ecol. Biogeogr.*, 20(1):73–86, 2011.

L. Knorr-Held and H. Rue. On block updating in Markov Random Field models for disease mapping. *Scand. J. Statist.*, 29(4):597–614, 2002.

F. Lindgren and H. Rue. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.*, 63(19):1–25, 2015. doi: 10.18637/jss.v063.i19.

F. Lindgren, R. Håvard, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc. Ser. B*, 73(4):423–498, 2011.

B. Matérn. Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelanden från statens Skogsforskningsinstitut 49, Statens Skogsforsningsinstitut, Stockholm, Sweden, 1960.

D. W. Nychka. Spatial-process estimates as smoothers. In M. G. A. Schimek, editor, *Smoothing and Regression: Approaches, Computation, and Application*, pages 393–424. Wiley, New York, USA, 2000.

C. J. Paciorek and J. S. McLachlan. Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Am. Statist. Assoc.*, 104(486):608–622, 2009.

R. Päivinen, M. Lehikoinen, A. Schuck, T. Häme, S. Väätäinen, P. Kennedy, and S. Folving. *Combining earth observation data and forest statistics*. EuroForIns, 2001.

B. Pirzamanbein, J. Lindström, A. Poska, S. Sugita, A.-K. Trondman, R. Fyfe, F. Mazier, A. B. Nielsen, J. O. Kaplan, A. E. Bjune, H. J. B. Birks, T. Giesecke, M. Kangur, M. Latałowa, L. Marquer, B. Smith, and M.-J. Gaillard. Creating spatially continuous maps of past land cover from point estimates: A new

statistical approach applied to pollen data. *Ecol. Complex.*, 20(0):127 – 141, 2014.

J. Pongratz, C. Reick, T. Raddatz, and M. Claussen. A reconstruction of global agricultural areas and land cover for the last millennium. *Global. Biogeochem. Cy.*, 22(3), 2008.

G. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4(4):337–358, 2003.

G. Roberts, A. Gelman, and W. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7:110–120, 1997.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. Roy. Statist. Soc. Ser. B*, 60(1):255–268, 1998.

J. S. Rosenthal. Optimal proposal distributions for adaptive MCMC. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 93–111. Chapman & Hall/CRC, 2011.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2004.

A. Schmidt, J. Hoeting, J. Batista Pereira, and P. Paulo Vieira. Mapping malaria in the amazon rain forest: a spatio-temporal mixture model. In *The Oxford Handbook of Applied Bayesian Statistics*, pages 90–117. Oxford University Press, 2010.

A. Schuck, J. van Brusselen, R. Päivinen, T. Häme, P. Kennedy, and S. Folving. Compilation of a calibrated European forest map derived from NOAA-AVHRR data. EFI Internal Report 13, EuroForIns, 2002.

D. Simpson, H. Rue, T. Martins, A. Riebler, and S. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. Technical Report 1403.4630v4, arXiv, 2015. URL `http://arxiv.org/abs/1403.4630v4`.

B. Smith, I. C. Prentice, and M. T. Sykes. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Global. Ecol. Biogeogr.*, 10(6):621–637, 2001.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

G. Strandberg, E. Kjellström, A. Poska, S. Wagner, M.-J. Gaillard, A.-K. Trondman, A. Mauri, B. A. S. Davis, J. O. Kaplan, H. J. B. Birks, A. E. Bjune, R. Fyfe, T. Giesecke, L. Kalnina, M. Kangur, W. O. van der Knaap, U. Kokfelt, P. Kuneš, M. Latałowa, L. Marquer, F. Mazier, A. B. Nielsen, B. Smith, H. Seppä, and S. Sugita. Regional climate model simulations for Europe at 6 and 0.2 k bp: sensitivity to changes in anthropogenic deforestation. *Clim. Past.*, 10(2):661–680, 2014.

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

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

H. Tjelmeland and K. V. Lund. Bayesian modelling of spatial compositional data. *J. Appl. Stat.*, 30(1):87–100, 2003.

A.-K. Trondman, M.-J. Gaillard, F. Mazier, S. Sugita, R. Fyfe, A. B. Nielsen, C. Twiddle, P. Barratt, H. J. B. Birks, A. E. Bjune, L. Björkman, A. Broström, C. Caseldine, R. David, J. Dodson, W. Dörfler, E. Fischer, B. van Geel, T. Giesecke, T. Hultberg, L. Kalnina, M. Kangur, P. van der Knaap, T. Koff, P. Kuneš, P. Lagerås, M. Latałowa, J. Lechterbeck, C. Leroyer, M. Leydet, M. Lindbladh, L. Marquer, F. J. G. Mitchell, B. V. Odgaard, S. M. Peglar, T. Persson, A. Poska, M. Rösch, H. Seppä, S. Veski, and L. Wick. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. *Glob. Change Biol.*, 21(2):676–697, 2015.

G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Comput.*, 2:5–16, 1981.

P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

C

Paper C

Analysing the sensitivity of pollen based land cover maps to different auxiliary variables

Behnaz Pirzamanbein[1,2], Anneli Poska[3,4], Johan Lindström[1]

[1] *Centre for Mathematical Sciences, Lund University, Sweden* [2] *Centre for Environmental and Climate Research, Lund University, Sweden* [3] *Department of Physical Geography and Ecosystems Analysis, Lund University, Sweden* [4] *Institute of Geology, Tallinn University of Technology, Estonia*

Abstract

In this paper, we aim to analyse the sensitivity of a spatial statistical model used to reconstruct past land-cover composition based on pollen data. The pollen data are irregularly places point observations, depicting the land cover composition of the area surrounding the studied sites. The statistical model is based on a mean and spatial dependence structure. The spatial dependency is modelled using Gaussian Markov Random Fields and the mean structure is a linear regression based on six different sets of covariates. The considered covariates include modern elevation, two different anthropogenic land-cover change scenarios, and two potential natural vegetation scenarios produced by a dynamic vegetation model forced with output from two different climate models. The estimation of the parameters and reconstruction of the land cover is done using Markov Chain Monte Carlo methods for three time periods, 1900 CE, 1725 CE, and 4000 BCE. The results are evaluated using deviance information criteria (DIC) and cross validation for six different models and all the time periods. For the recent time periods we compared the land-cover reconstruction based on pollen data and different covariates with present day European forest map.

According to the conducted statistical tests the model produced well comparable results despite considerable differences in applied auxiliary data. This implies

that the developed Bayesian hierarchical model is a robust spatial interpolation tool with high capacity to un-distortedly transmit the information provided by pollen based data and low dependence on the choice of covariates. However, usage of auxiliary data with high spatial detail improves the model performance for the areas with low observational data coverage.

Key words: Spatial interpolation, Pollen based vegetation reconstruction, Gaussian Markov random field, Sensitivity study, Past land cover, Anthropogenic land-cover changes.

1 Introduction

The importance of terrestrial land cover for the global carbon cycle and its impact on the climate system is well recognized (e.g. Arneth et al., 2010, Brovkin et al., 2006, Christidis et al., 2013, Claussen et al., 2001). Many studies have found large climatic effects associated with changes in land cover. Forecast simulations evaluating the effects of human induced global warming predict a considerable amplification of future climate change for Arctic areas. The amplification is, due to a number of biogeophysical and -chemical feedbacks brought by the northward advancement of boreal shrub and treeline (Chapman and Walsh, 2007, Koenigk et al., 2013, Miller and Smith, 2012, Richter-Menge et al., 2011, Zhang et al., 2013). The past anthropogenic deforestation of the temperate zone in Europe was lately demonstrated to have an impact on regional climate similar in amplitude to present day climate change (Strandberg et al., 2014). However, studies on the effects of vegetation and land-use changes on past climate and carbon cycle often report considerable differences and uncertainties in their model predictions (de Noblet-Ducoudré et al., 2012, Olofsson, 2013).

One of the reasons for such widely diverging results could be the differences in past land-cover descriptions used by climate modellers. Possible land-cover descriptions range from static present-day land cover (Strandberg et al., 2011), over simulated potential natural land-cover from dynamic (or static) vegetation models (DVMs) (e.g. Brovkin et al., 2002, Hickler et al., 2012), to past land-cover scenarios combining DVM derived potential vegetation with estimates of anthropogenic land-cover change (ALCC) (de Noblet-Ducoudré et al., 2012, Pongratz et al., 2008, Strandberg et al., 2014). Differences in input climates, inherent mechanistic and parametrisation differences of DVMs (Prentice et al., 2007, Scheiter et al., 2013), and significant variation in land-use estimates between the

existing ALCC scenarios (e.g. Gaillard et al., 2010, Kaplan et al., 2009, Klein Goldewijk et al., 2011, Pongratz et al., 2008) further contribute to the differences in past land-cover descriptions. These differences can lead to largely diverging estimates of past land-cover dynamics even when the most advanced models are used.

The palaeoecological observation based land-cover reconstructions (LCR) recently published by Pirzamanbein et al. (2014, 2015) were designed to overcome the above described problems. And to provide an alternative, observation based, land-cover description applicable for a range of studies on past vegetation and its interactions with climate, soil and humans. These reconstructions use the pollen based land-cover composition (PbLCC) published by Trondman et al. (2015) as a source of information on past land-cover composition. The PbLCC are point estimates, depicting the land-cover composition of the area surrounding the studied sites. To fill the gaps between these observations and to acquire a spatially continuous land-cover reconstruction, spatial interpolation is necessary. Conventional interpolation methods might struggle when handling noisy, spatially heterogeneous data (De Knegt et al., 2010, Heuvelink et al., 1999), but alternative statistical methods for handling spatially structured data exist (e.g. Blangiardo and Cameletti, 2015, Gelfand et al., 2010).

In Pirzamanbein et al. (2015) a statistical model based on Gaussian Markov Random Fields (GMRFs, Lindgren et al., 2011, Rue and Held, 2004) was developed to provide a reliable, computationally effective and freeware based spatial interpolation technique. The current study aims at determining the robustness of the model. To evaluate its capacity to un-distortedly recover information provided by PbLCC observations on past vegetation composition, and to analyse the models sensitivity to auxiliary datasets.

2 Material and methods

The studied area covers temperate, boreal and alpine-arctic biomes of central and northern Europe (45°N to 71°N and 10°W to 30°E). The Pollen based land-cover composition (PbLCC) published in (Trondman et al., 2015) consists of proportions of coniferous forest, broadleaved forest and un-forested land presented as gridded ($1^\circ \times 1^\circ$) data points placed irregularly across northern-central Europe. Altogether 175 grid cells containing the observational data were available for 1900 CE, 181 for 1725 CE and 196 for the 4000 BCE time-period (Figure 1,

column 2).

Four different model derived datasets depicting past land-cover were considered as potential (auxiliary) datasets:

K-L_{RCA3}: Combines the ALCC scenario KK10 (Kaplan et al., 2009) and potential natural vegetation (PNV) composition estimated by the dynamic vegetation model (DVM) LPJ-GUESS (Lund-Potsdam-Jena General Ecosystem Simulator; Sitch et al., 2003, Smith et al., 2001). Climate forcing for the DVM was derived from RCA3 (Rossby Centre Regional Climate Model, Samuelsson et al., 2011) at annual time and $0.44° \times 0.44°$ spatial resolution (Figure 1, column 3),

K-L_{ESM}: Combines the ALCC scenario KK10 and the PNV composition from LPJ-GUESS. For this dataset, the climate forcing for the DVM was derived from the Earth System Model (ESM; Mikolajewicz et al., 2007) at centennial time and $5.6° \times 5.6°$ spatial resolution. To interpolate data into annual time and $0.5° \times 0.5°$ spatial resolution climate data from 1901–1930 CE provided by the Climate Research Unit (CRU) was used (Figure 1, column 4),

H-L_{RCA3}: Combines the ALCC scenario from the History Database of the Global Environment (HYDE; Klein Goldewijk et al., 2011) and the PVN composition from LPJ-GUESS with RCA3 climate forcing (Figure 1, column 5),

H-L_{ESM}: Combines the ALCC scenario from HYDE and the PVN composition from LPJ-GUESS with ESM climate forcing (Figure 1, column 6).

In addition, elevation data used in modelling was obtained from the Shuttle Radar Topography Mission ($SRTM_{elev}$, Becker et al., 2009) (Figure 1, column 1 row 2).

Finally, a modern forest map based on data from the European Forest Institute (EFI) is used for evaluation of the model's performance for the 1900 CE time period. The EFI forest map (EFI-FM) is based on a combination of satellite data (NOAA-AVHRR) and national forest-inventory statistics from 1990—2005 (Päivinen et al., 2001, Schuck et al., 2002) (Figure 1, column 1 row 1).

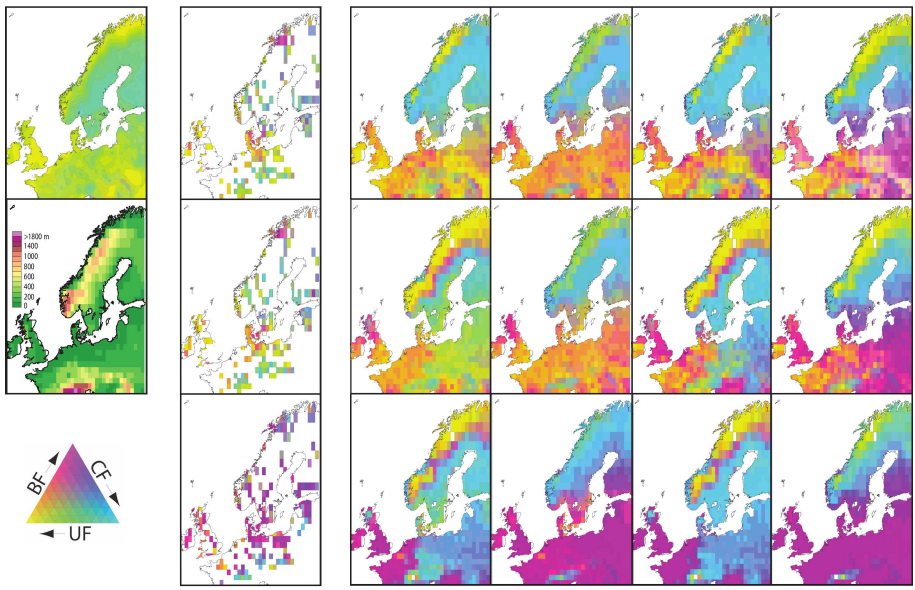


Figure 1: Data used in the modelling. The first column shows (from top to bottom) the EFI forest map, $\text{SRTM}_{\text{elev}}$, and the colorkey for the compositional data. The remaining columns gives (from left to right) the pollen based land-cover composition (PbLCC, Trondman et al., 2015) and the four model based compositions that could be used as covariates: K-L_{RCA3}, K-L_{ESM}, H-L_{RCA3}, and H-L_{ESM}; with the three rows representing (from top to bottom) the time periods 1900 CE, 1725 CE, and 4000 BCE.

2.1 Statistical model for land-cover compositions

A Bayesian hierarchical model (Figure 2) is used to model the PbLCC data. For each component of PbLLC, we assume an underlying compositional vector describing the proportions of land cover; coniferous forest, broadleaved forest and un-forested land. The effect of covariates and spatial structure are incorporated in the underlying compositional vector.

To account for observational uncertainty in the compositions, the PbLCC are modelled as draws from a Dirichlet distribution given concentrated parameter α (controlling the uncertainty) and the vector of proportions $\mathbf{Z}$,

$$\mathbf{Y}_{\text{PbLCC}}|\mathbf{Z}, \alpha \sim \mathsf{Dir}(\alpha\mathbf{Z}) \qquad \alpha > 0,\ \mathbf{Z}_k \in (0,1),\ \sum_k \mathbf{Z}_k = 1.$$

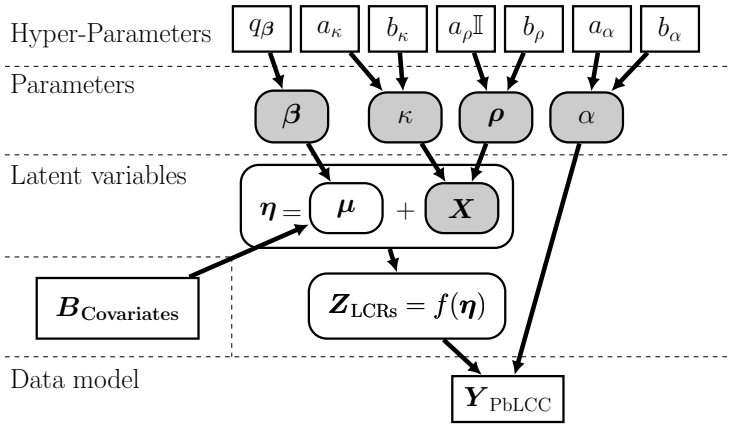


Figure 2: Hierarchical graph describing the conditional dependencies between the model inputs (white rectangle) and parameters (gray rounded rectangle) which need to be estimated. The white rounded rectangles are computed based on the estimations.

To account for the spatial dependence in the proportions, $\mathbf{Z}$ is modelled as a transformation, f, of a latent GMRF, $\boldsymbol{\eta}$:

$$\mathbf{Z} = f(\boldsymbol{\eta}) \qquad f : \mathbb{R}^2 \to (0,1)^3$$

$$\mathbf{Z}_k = \begin{cases} \dfrac{\exp(\boldsymbol{\eta}_k)}{1 + \sum_{i=1}^{2} \exp(\boldsymbol{\eta}_i)} & \text{for } k = 1, 2 \\ \dfrac{1}{1 + \sum_{i=1}^{2} \exp(\boldsymbol{\eta}_i)} & \text{for } k = 3 \end{cases}.$$

The inverse of f is called the additive log-ratio transformation (alr, Aitchison, 1986), i.e. $\boldsymbol{\eta}_k = \log\left(\mathbf{Z}_k / \mathbf{Z}_3\right)$, $k = 1, 2$. The alr transformation can been seen as the multivariate extension of a logit transformation.

The latent field is modelled with a mean structure $\boldsymbol{\mu}$ and a spatially dependent residual $\mathbf{X}$,

$$\boldsymbol{\eta} = \mathbf{X} + \boldsymbol{\mu}$$

where $\mathbf{X}$ is GMRF with a separable covariance structure;

$$\mathbf{X} | \kappa, \boldsymbol{\rho} \sim \mathrm{N}(0, \boldsymbol{\rho}^{-1} \otimes \mathbf{Q}(\kappa))$$

where $\mathbf{Q}(\kappa)$ is the precision matrix of GMRF, κ is the scale parameter which controls the range of spatial dependency and $\boldsymbol{\rho}$ controls the variation within and between the fields $\mathbf{X}$ (See Pirzamanbein et al., 2015, for details).

The mean structure is modelled as a linear regression $\boldsymbol{\mu} = \mathbf{B}\boldsymbol{\beta}$, i.e. a combination of covariates $\mathbf{B}$ and regression coefficients $\boldsymbol{\beta}$. The main focus of this paper is to evaluate the model sensitivity to the choice of covariates. The PbLCC is modelled based on six different sets of covariates (Figure 1): 1) Intercept, 2) $\text{SRTM}_{\text{elev}}$, 3) K-L_{ESM}, 4) K-L_{RCA3}, 5) H-L_{ESM}, and 6) H-L_{RCA3}. Table 1 shows the different models and the corresponding covariates included in the model.

Model	Covariates					
	Intercept	$\text{SRTM}_{\text{elev}}$	K-L_{ESM}	K-L_{RCA3}	H-L_{ESM}	H-L_{RCA3}
Constant	x					
Elevation	x	x				
K-L_{ESM}	x	x	x			
K-L_{RCA3}	x	x		x		
H-L_{ESM}	x	x			x	
H-L_{RCA3}	x	x				x

Table 1: Six different models and corresponding covariates.

The model description is completed by specifying prior distributions for the model parameters. Wide but proper priors are assigned for $\alpha, \kappa, \boldsymbol{\rho}$ and β. Specifically, a Gamma prior is chosen for the uncertainty and scale parameters, α and κ, i.e. $\Gamma(1.5, 0.1)$ and $\Gamma(1.5, 0.1)$. A Gaussian prior for the regression parameters $\boldsymbol{\beta}$, with zero expectation and small precision $q_\beta = 10^{-3}$. The $\boldsymbol{\rho}$ is assigned an inverse Wishart prior, $IW(\mathbb{I}, 10)$, where $\mathbb{I}$ is a 2×2 identity matrix.

2.2 Inference and associated uncertainties

The Markov Chain Monte Carlo (MCMC) method is used to estimate the parameters and to reconstruct the land-cover composition, $\mathbf{Z}_{\text{LCRs}}$, with 100 000 MCMC samples and a burn-in sample size of 10 000. First, the algorithm updates $\mathbf{X}$, $\boldsymbol{\beta}$ and α together given PbLCC data ($\mathbf{Y}_{\text{PbLCC}}$), κ, and ρ. Using the updated $\mathbf{X}$, the parameters of the GMRF, κ and $\boldsymbol{\rho}$, are updated. Details of the MCMC implementation can be found in Pirzamanbein et al. (2015).

In each MCMC iteration, the samples of $\boldsymbol{\eta}$ are obtained by adding the spatial

dependency field $\mathbf{X}$ and the effect of covariates through $\mathbf{B}\boldsymbol{\beta}$. Applying the alr transformation to the $\boldsymbol{\eta}$ samples, MCMC samples for $\mathbf{Z}$ are obtained. The land-cover reconstruction is then computed as $\mathbf{Z}_{\text{LCRs}} = \text{E}(\mathbf{Z}|\mathbf{Y}_{\text{PbLCC}})$, by averaging the MCMC samples.

The uncertainties of the land-cover reconstruction $\text{V}(\mathbf{Z}|\mathbf{Y}_{\text{PbLCC}})$ are assessed by constructing predictive regions (PR) using the MCMC samples at each location. The predictive region is constructed to measure the uncertainty associated with the data given the data model parameter α and the underlying fields $\mathbf{Z}$. For predictive regions, we use the $\mathbf{Z}$ and corresponding α from each MCMC sample and sample new Dirichlet observations. These new Dirichlet draws are then transformed to $\mathbb{R}^2$. Then an elliptical predictive region containing 95% of the MCMC samples is constructed (Figure 3 left plot). Thereafter the elliptical predictive region is transformed from $\mathbb{R}^2$ to a ternary predictive region, $(0, 1)^3$ (Figure 3 right plot). In order to compare the uncertainties of different model land-cover reconstructions, we report the fraction of the unit triangle covered by the ternary PR. This is done by distributing points in the ternary diagram and computing the fraction as the number of points laying inside the PR divided by total number of points in the ternary triangle.

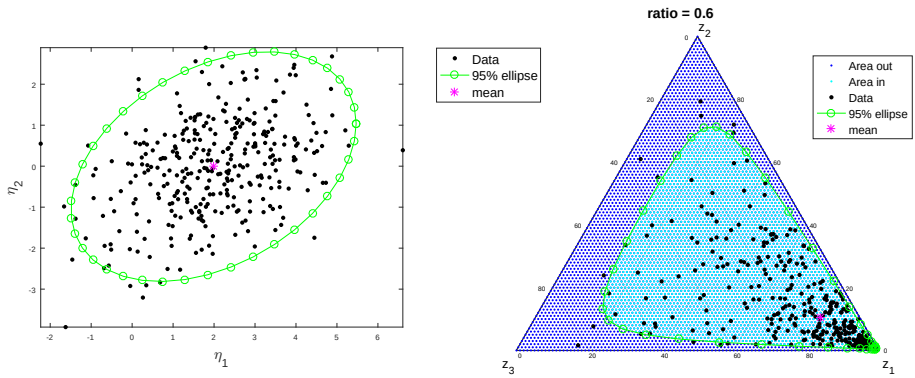


Figure 3: The left plot shows the 95% elliptical predictive region. The right ternary diagram shows the transformed 95% predictive region together with the corresponding fraction, 60%, compared to the whole triangle.

2.3 Testing the model performance

To evaluate the model performance, we compared the land-cover reconstructions from different models for 1900 CE time period with the European Forest Institute forest map (EFI-FM) by computing the average compositional distances (ACD). The compositional distances between two different compositions, **U** and **V**, are computed using

$$\Delta(\mathbf{U}-\mathbf{V}) = \Delta(\mathbf{u}-\mathbf{v}) = \Big((\mathbf{u}-\mathbf{v})^{\top}\mathbf{H}^{-1}(\mathbf{u}-\mathbf{v})\Big)^{1/2}$$

where $\mathbf{u} = alr(\mathbf{U})$, $\mathbf{v} = alr(\mathbf{V})$ and $\mathbf{H}$ is a 2×2 matrix, neutralizing the choice of denominator in the alr transformation, with elements $H_{ij} = 2$ if $i = j$, and $H_{ij} = 1$ if $i \neq j$. These distances are then averaged over all locations. This measure is similar to root mean square error in $\mathbb{R}^2$ but it accounts for compositional properties, i.e. each component of the compositions is between $(0, 1)$ and sum of all the components are 1.

Since no independent observational data exists for the 1725 CE and 4000 BCE time periods, we applied a 6-fold cross-validation scheme (Friedman et al., 2001, Ch. 7.10) for all the six models and three time periods. The PbLCC data are divided into 6 randomly selected groups and in each round the distance between predictions for group l given the rest of the data, $\mathsf{E}(\mathbf{Z}_l|\mathbf{Y}_{\mathrm{PbLCC},k}\ k \notin l)$, and left out data, $\mathbf{Y}_{\mathrm{PbLCC},l}$, are computed.

To compare the predictive performance of the models, the Deviance Information Criteria (DIC; see Gelman et al., 2014, Ch. 7.2) is also computed for all models and time periods.

3 Results and discussion

A number of auxiliary datasets, including modern elevation, and four different model based estimates of the land cover for every corresponding time period (Figure 1) were used to compile the covariate datasets used in different models. Differences in land-cover estimates between the studied time-period are in general larger than the differences within a time period. However, the variation in extent of coniferous and broadleaved forests, and unforested areas inside any of the studied time periods is considerable. These substantial variations illustrate the large deviances between the model based estimates of past land-cover composition caused by differences in climate forcing and ALCC scenarios. Considerable

variability in climate model simulations and ALCC scenarios is well recognized (e.g. Gaillard et al., 2010, Gladstone et al., 2005, Harrison et al., 2014). The effects of the differences in climate forcing on land-cover estimates presented here are especially pronounced for central and eastern Europe, and for elevated areas in western and northern Scandinavia and the Alps. The differences are clearly discernible for all considered time-periods. The variance in applied anthropogenic deforestation scenarios is especially pronounced for western Europe during the 1725 and 1900 CE time periods. The importance of reliable land-cover representation for studies on biogeophysical impacts of anthropogenic land-cover change is well recognized by the climate modelling community (Pitman et al., 2009, Strandberg et al., 2014) and usage of the above described, solely model based land-cover representations, to assess the impact of the past anthropogenic changes on climate and terrestrial nutrient cycles leads to largely diverging results.

The impact of the above described auxiliary data on the statistical model's performance was assessed by comparing the land-cover reconstructions produced by six statistical models employing different covariate datasets (Table 1).

To illustrate the structure of the statistical model, step by step advancement from auxiliary data (model derived land-cover) to final statistical estimates, for 1725 CE, are given in Figures 4. Figure 4 shows, for two locations, how considerable differences in K-LRCA3 and K-LESM are reduced by scaling with the regression coefficients, β. The two land cover estimates are then further subject to similar adjustments due to intercept and $SRTM_{elev}$, and finally similar spatial dependent effects. Corresponding progressions for continental maps of land-cover are given in Figure 5 for constant, Elevation and K-L_{RCA3} model for 1900 CE .

The final land-cover reconstructions achieved by fitting the model to observed PbLCC are very similar between all considered datasets. While, in general, the land-cover reconstructions produced by different models are very similar, the model performance for the areas with low observational data coverage (e.g. eastern and south-eastern Europe) is considerably improved by including covariates based on auxiliary data exhibiting distinct spatial structures for the given areas (Figure 6).

The resulting land-cover reconstructions exhibit considerably higher similarity than the auxiliary land-cover datasets for all tested models and time-windows (Figure 6). The predictive regions indicate the capability of all the models in capturing the PbLCC data and shows similar reconstruction uncertainties (Figure 7).

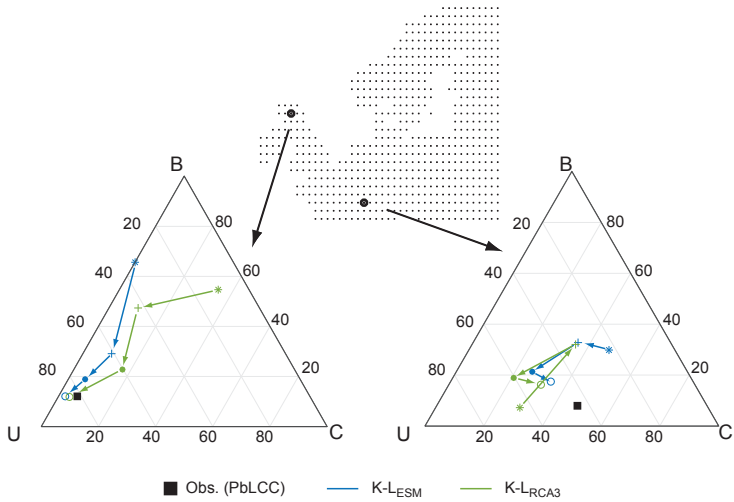


Figure 4: Advancement of the model for two locations in 1725 CE. Starting from the value of the K-L$_{RCA3}$ and K-L$_{ESM}$ covariates (∗), the cumulative effects of regression coefficients, β, (+); the intercept and SRTM$_{elev}$ covariates (●); and, finally, the spatial dependency structures (○), are illustrated. With the final points (○) corresponding to the land-cover reconstructions, $\mathbf{Z}_{LCRs}$, and ■ marking the observed pollen based land-cover composition.

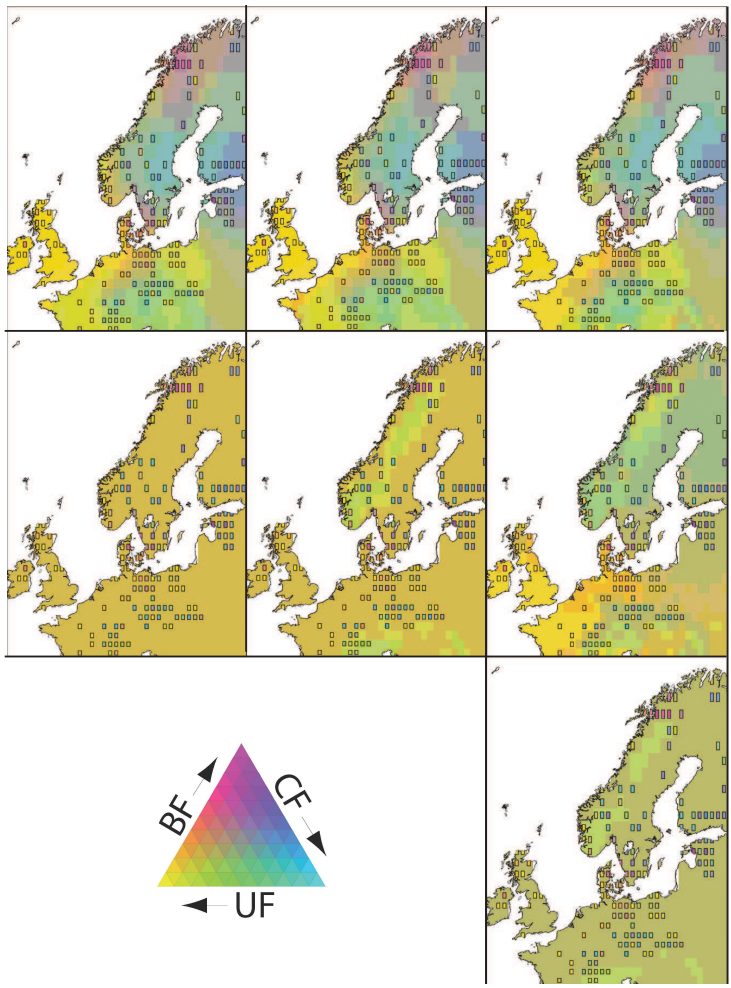


Figure 5: Advancement of three different models for the 1900 CE time period, from left to right the models are (see Table 1): Constant, Elevation, K-L_{RCA3}. The bottom row shows the effect of intercept and $SRTM_{elev}$. The second row shows the mean structure, $\boldsymbol{\mu}$, for each model. Finally, the top row shows the resulting land-cover reconstructions, $\mathbf{Z}_{LCRs}$, obtained by adding the spatial dependency structure to the mean structure.

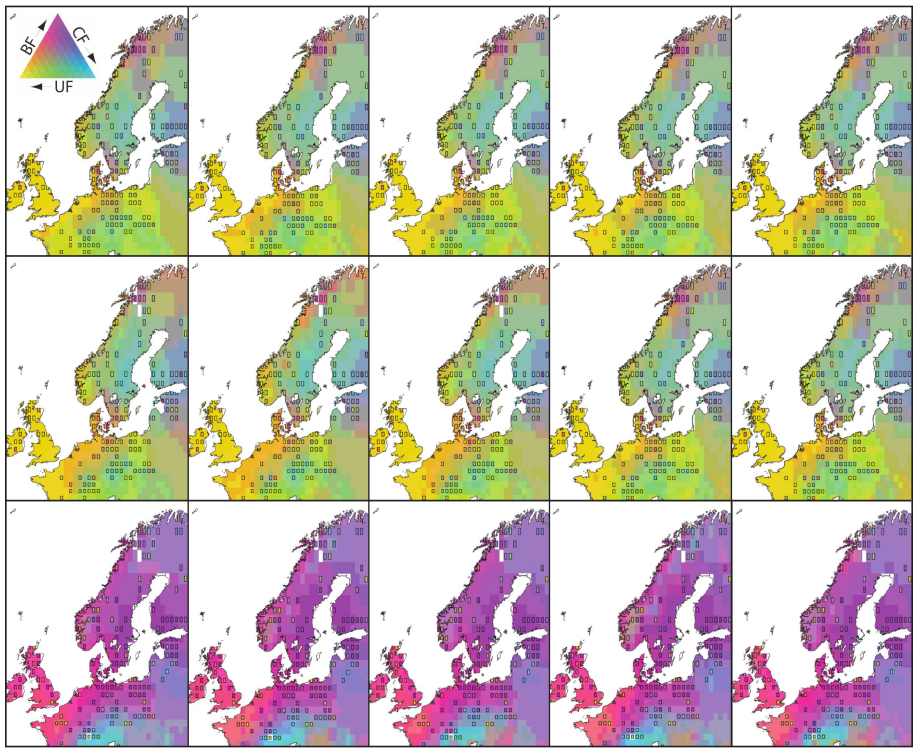


Figure 6: Land-cover reconstructions using pollen based land-cover compositions (PbLCC) for the 1950 CE, 1750 CE and 4000 BCE time periods. The reconstructions are based on six different models (Table 1) with different auxiliary datasets. Here reconstructions for Elevation, K-L_{RCA3}, K-L_{ESM}, H-L_{RCA3}, H-L_{ESM} are shown. Locations and compositional values of the available PbLCC data are given by the black rectangles.

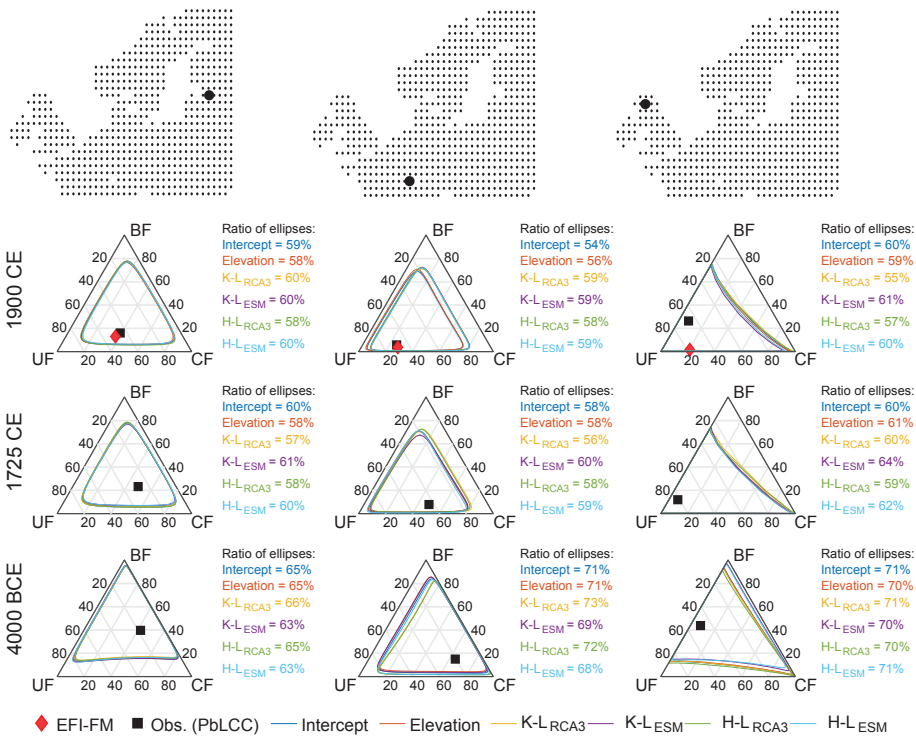


Figure 7: The prediction regions and fraction of the ternary triangle covered by these regions are presented for three locations, the six models (Table 1), and the 1950 CE, 1750 CE and 4000 BCE time periods.

The ACDs computed between the land-cover reconstructions and the EFI-FM for 1900 CE for all models show practically identical (1.47 to 1.48) distances between the reconstructions and the EFI-FM, and small differences among the six presented models (Table. 2). Furthermore, the DIC results show no advantage among the six tested models for the different time periods (Table. 3), and 6-fold cross validation results for all time periods implies that no clear preference can be given to any of the models (Table. 4). These results clearly suggest that the developed model is robust to the choice of covariates and well applicable for interpolating the land-cover composition represented by pollen based observations with irregular spatial distribution.

ACD

	1900 CE					
Model	EFI-FM	Elevation	K-L_{RCA3}	K-L_{ESM}	H-L_{RCA3}	H-L_{ESM}
Constant	1.47	0.06	0.19	0.17	0.19	0.18
Elevation	1.48		0.18	0.16	0.18	0.17
K-L_{RCA3}	1.47			0.08	0.06	0.11
K-L_{ESM}	1.47				0.10	0.07
H-L_{RCA3}	1.47					0.08
H-L_{ESM}	1.47					
	1725 CE					
Constant		0.11	0.19	0.16	0.19	0.18
Elevation			0.12	0.14	0.14	0.16
K-L_{RCA3}				0.15	0.08	0.18
K-L_{ESM}					0.16	0.07
H-L_{RCA3}						0.17
	4000 BCE					
Constant		0.12	0.19	0.21	0.21	0.23
Elevation			0.12	0.19	0.16	0.21
K-L_{RCA3}				0.19	0.07	0.19
K-L_{ESM}					0.21	0.07
H-L_{RCA3}						0.21

Table 2: The average compositional distances among the six models fitted to the data for each of the three time periods.

DIC	1900 CE	1725 CE	4000 BCE
Constant	**-562**	-656	-591
Elevation	-557	-668	-590
K-L_{RCA3}	-559	**-673**	-588
K-L_{ESM}	-551	-654	-601
H-L_{RCA3}	-559	-672	-594
H-L_{ESM}	-554	-654	**-607**

Table 3: Deviance information criteria (DIC) for each of the models and time periods.

ACD	1900 CE	1725 CE	4000 BCE
Constant	0.98	1.13	1.19
Elevation	0.98	1.11	1.20
K-L_{RCA3}	0.99	1.13	1.18
K-L_{ESM}	0.99	1.12	1.18
H-L_{RCA3}	**0.97**	**0.97**	1.17
H-L_{ESM}	1.00	1.12	**1.17**

Table 4: Average compositional distances from 6-fold cross-validations for each of the models, and time periods.

4 Conclusion

The performance of the statistical model, explained in Section 2, to reconstruct the pollen based past land cover was tested in order to analyse its sensitivity to auxiliary datasets.

The considered auxiliary datasets were complied using most commonly utilized sources of the spatially explicit past land cover data (estimates produced by a dynamic vegetation model and anthropogenic land cover changes scenarios). These datasets exhibit considerable model and/or input dependant differences in their recreation of past land cover. Emphasizing the need for the independent and observation based past land cover maps created in this paper.

The model sensitivity to usage of different auxiliary datasets was validated by calculating deviance information criteria (DIC) and using cross validation for all the time periods. For the recent time period, 1900 CE, the land-cover reconstructions from the different models were also compared against a present day forest map.

The evaluation indicates that the applied statistical model is robust and well applicable for interpolating the pollen based land-cover composition with irregular spatial distribution. The spatial resolution of the covariates improves the interpolation results for areas with low observational data coverage, however the overall performance remains unchanged.

Acknowledgement

The research presented in this paper is a contribution to the two Swedish strategic research areas Biodiversity and Ecosystems in a Changing Climate (BECC), and ModElling the Regional and Global Earth system (MERGE).

References

J. Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., 1986.

A. Arneth, S. P. Harrison, S. Zaehle, K. Tsigaridis, S. Menon, P. J. Bartlein, J. Feichter, A. Korhola, M. Kulmala, D. O'donnell, et al. Terrestrial biogeochemical feedbacks in the climate system. *Nat. Geo. Sci.*, 3(8):525–532, 2010.

J. Becker, D. Sandwell, W. Smith, J. Braud, B. Binder, J. Depner, D. Fabre, J. Factor, S. Ingalls, S. Kim, et al. Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Marine Geodesy*, 32(4):355–371, 2009.

M. Blangiardo and M. Cameletti. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley, 2015.

V. Brovkin, J. Bendtsen, M. Claussen, A. Ganopolski, C. Kubatzki, V. Petoukhov, and A. Andreev. Carbon cycle, vegetation, and climate dynamics in the Holocene: Experiments with the CLIMBER-2 model. *Global. Biogeochem. Cy.*, 16 (4):1139, 2002.

V. Brovkin, M. Claussen, E. Driesschaert, T. Fichefet, D. Kicklighter, M. Loutre, H. Matthews, N. Ramankutty, M. Schaeffer, and A. Sokolov. Biogeophysical effects of historical land cover changes simulated by six earth system models of intermediate complexity. *Clim. Dynam.*, 26(6):587–600, 2006.

W. L. Chapman and J. E. Walsh. Simulations of Arctic temperature and pressure by global coupled models. *J. Climate*, 20(4):609–632, 2007.

N. Christidis, P. A. Stott, G. C. Hegerl, and R. A. Betts. The role of land use change in the recent warming of daily extreme temperatures. *Geophys. Res. Lett.*, 40(3):589–594, 2013.

M. Claussen, V. Brovkin, and A. Ganopolski. Biogeophysical versus biogeochemical feedbacks of large-scale land cover change. *Geophys. Res. Lett.*, 28(6):1011–1014, 2001.

H. De Knegt, F. v. van Langevelde, M. Coughenour, A. Skidmore, W. De Boer, I. Heitkönig, N. Knox, R. Slotow, C. Van der Waal, and H. Prins. Spatial

autocorrelation and the scaling of species–environment relationships. *Ecology*, 91(8):2455–2465, 2010.

N. de Noblet-Ducoudré, J.-P. Boisier, A. Pitman, G. Bonan, V. Brovkin, F. Cruz, C. Delire, V. Gayler, B. van den Hurk, P. Lawrence, M. K. van der Molen, C. Müller, C. H. Reick, B. J. Strengers, , and A. Voldoire. Determining robust impacts of land-use-induced land cover changes on surface climate over North America and Eurasia: results from the first set of LUCID experiments. *J. Climate*, 25(9):3261–3281, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

M.-J. Gaillard, S. Sugita, F. Mazier, A.-K. Trondman, A. Brostrom, T. Hickler, J. O. Kaplan, E. Kjellström, U. Kokfelt, P. Kuneš, , C. Lemmen, P. Miller, J. Olofsson, A. Poska, M. Rundgren, B. Smith, G. Strandberg, R. Fyfe, A. Nielsen, T. Alenius, L. Balakauskas, L. Barnekov, H. Birks, A. Bjune, L. Björkman, T. Giesecke, K. Hjelle, L. Kalnina, M. Kangur, W. van der Knaap, T. Koff, P. Lagerås, M. Latałowa, M. Leydet, J. Lechterbeck, M. Lindbladh, B. Odgaard, S. Peglar, U. Segerström, H. von Stedingk, and H. Seppä. Holocene land-cover reconstructions for studies on land cover-climate feedbacks. *Clim. Past.*, 6:483–499, 2010.

A. Gelfand, P. J. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC Press, 2010.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2014.

R. M. Gladstone, I. Ross, P. J. Valdes, A. Abe-Ouchi, P. Braconnot, S. Brewer, M. Kageyama, A. Kitoh, A. Legrande, O. Marti, R. Ohgaito, B. Otto-Bliesner, W. R. Peltier, and G. Vettoretti. Mid-Holocene NAO: A pmip2 model intercomparison. *Geophys. Res. Lett.*, 32(16), 2005.

S. Harrison, P. Bartlein, S. Brewer, I. Prentice, M. Boyd, I. Hessler, K. Holmgren, K. Izumi, and K. Willis. Climate model benchmarking with glacial and mid-Holocene climates. *Clim. Dynam.*, 43(3–4):671–688, 2014.

G. Heuvelink et al. Propagation of error in spatial modelling with GIS. 1:207–217, 1999.

T. Hickler, K. Vohland, J. Feehan, P. A. Miller, B. Smith, L. Costa, T. Giesecke, S. Fronzek, T. R. Carter, W. Cramer, I. Kühn, and M. T. Sykes. Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global. Ecol. Biogeogr.*, 21(1):50–63, 2012.

J. O. Kaplan, K. M. Krumhardt, and N. Zimmermann. The prehistoric and preindustrial deforestation of Europe. *Quaternary. Sci. Rev.*, 28(27):3016–3034, 2009.

K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Global. Ecol. Biogeogr.*, 20(1):73–86, 2011.

T. Koenigk, L. Brodeau, R. G. Graversen, J. Karlsson, G. Svensson, M. Tjernström, U. Willén, and K. Wyser. Arctic climate change in 21st century CMIP5 simulations with EC-Earth. *Clim. Dynam.*, 40(11-12):2719–2743, 2013.

F. Lindgren, R. Håvard, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc. Ser. B*, 73(4):423–498, 2011.

U. Mikolajewicz, M. Gröger, E. Maier-Reimer, G. Schurgers, M. Vizcaíno, and A. M. Winguth. Long-term effects of anthropogenic co2 emissions simulated with a complex earth system model. *Clim. Dynam.*, 28(6):599–633, 2007.

P. A. Miller and B. Smith. Modelling tundra vegetation response to recent arctic warming. *Ambio*, 41(3):281–291, 2012.

J. Olofsson. *The Earth: climate and anthropogenic interactions in a long time perspective*. Lund University, 2013.

R. Päivinen, M. Lehikoinen, A. Schuck, T. Häme, S. Väätäinen, P. Kennedy, and S. Folving. *Combining earth observation data and forest statistics*. EuroForIns, 2001.

B. Pirzamanbein, J. Lindström, A. Poska, S. Sugita, A.-K. Trondman, R. Fyfe, F. Mazier, A. B. Nielsen, J. O. Kaplan, A. E. Bjune, H. J. B. Birks, T. Giesecke, M. Kangur, M. Latałowa, L. Marquer, B. Smith, and M.-J. Gaillard. Creating spatially continuous maps of past land cover from point estimates: A new

statistical approach applied to pollen data. *Ecol. Complex.*, 20(0):127 – 141, 2014.

B. Pirzamanbein, J. Lindström, A. Poska, and M.-J. Gaillard. Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *arXiv preprint arXiv:1511.06417*, 2015.

A. Pitman, N. de Noblet-Ducoudré, F. Cruz, E. Davin, G. Bonan, V. Brovkin, M. Claussen, C. Delire, L. Ganzeveld, V. Gayler, B. J. J. M. van den Hurk, P. J. Lawrence, M. K. van der Molen, C. Müller, C. H. Reick, S. I. Seneviratne, B. J. Strengers, and A. Voldoire. Uncertainties in climate responses to past land cover change: First results from the LUCID intercomparison study. *Geophys. Res. Lett.*, 36(14), 2009.

J. Pongratz, C. Reick, T. Raddatz, and M. Claussen. A reconstruction of global agricultural areas and land cover for the last millennium. *Global. Biogeochem. Cy.*, 22(3), 2008.

I. C. Prentice, A. Bondeau, W. Cramer, S. P. Harrison, T. Hickler, W. Lucht, S. Sitch, B. Smith, and M. T. Sykes. Dynamic global vegetation modeling: quantifying terrestrial ecosystem responses to large-scale environmental change. In *Terrestrial ecosystems in a changing world*, pages 175–192. Springer, 2007.

J. A. Richter-Menge, M. O. Jeffries, and J. E. Overland. *Arctic report card 2011*. National Oceanic and Atmospheric Administration, 2011.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2004.

P. Samuelsson, C. G. Jones, U. Willén, A. Ullerstig, S. Gollvik, U. Hansson, C. Jansson, E. Kjellström, G. Nikulin, and K. Wyser. The rossby centre regional climate model rca3: model description and performance. *Tellus A*, 63 (1):4–23, 2011.

S. Scheiter, L. Langan, and S. I. Higgins. Next-generation dynamic global vegetation models: learning from community ecology. 198(3):957–969, 2013.

A. Schuck, J. van Brusselen, R. Päivinen, T. Häme, P. Kennedy, and S. Folving. Compilation of a calibrated European forest map derived from NOAA-AVHRR data. EFI Internal Report 13, EuroForIns, 2002.

S. Sitch, B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. Kaplan, S. Levis, W. Lucht, M. Sykes, K. Thonicke, and S. Venevsky. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Glob. Change Biol.*, 9(2):161–185, 2003.

B. Smith, I. C. Prentice, and M. T. Sykes. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Global. Ecol. Biogeogr.*, 10(6):621–637, 2001.

G. Strandberg, J. Brandefelt, E. Kjellström, and B. Smith. High-resolution regional simulation of last glacial maximum climate in Europe. *Tellus. A*, 63(1): 107–125, 2011.

G. Strandberg, E. Kjellström, A. Poska, S. Wagner, M.-J. Gaillard, A.-K. Trondman, A. Mauri, B. A. S. Davis, J. O. Kaplan, H. J. B. Birks, A. E. Bjune, R. Fyfe, T. Giesecke, L. Kalnina, M. Kangur, W. O. van der Knaap, U. Kokfelt, P. Kuneš, M. Latałowa, L. Marquer, F. Mazier, A. B. Nielsen, B. Smith, H. Seppä, and S. Sugita. Regional climate model simulations for Europe at 6 and 0.2 k bp: sensitivity to changes in anthropogenic deforestation. *Clim. Past.*, 10(2):661–680, 2014.

A.-K. Trondman, M.-J. Gaillard, F. Mazier, S. Sugita, R. Fyfe, A. B. Nielsen, C. Twiddle, P. Barratt, H. J. B. Birks, A. E. Bjune, L. Björkman, A. Broström, C. Caseldine, R. David, J. Dodson, W. Dörfler, E. Fischer, B. van Geel, T. Giesecke, T. Hultberg, L. Kalnina, M. Kangur, P. van der Knaap, T. Koff, P. Kuneš, P. Lagerås, M. Latałowa, J. Lechterbeck, C. Leroyer, M. Leydet, M. Lindbladh, L. Marquer, F. J. G. Mitchell, B. V. Odgaard, S. M. Peglar, T. Persson, A. Poska, M. Rösch, H. Seppä, S. Veski, and L. Wick. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. *Glob. Change Biol.*, 21(2):676–697, 2015.

W. Zhang, P. A. Miller, B. Smith, R. Wania, T. Koenigk, and R. Döscher. Tundra shrubification and tree-line advance amplify arctic climate warming: results from an individual-based dynamic vegetation model. 8(3):034023, 2013.

D

Paper D

Reconstruction of Past Human Land Use from Pollen Data and Anthropogenic Land Cover Changes Scenarios

Behnaz Pirzamanbein[1,2], Johan Lindström[1]

[1] *Centre for Mathematical Sciences, Lund University, Sweden* [2] *Centre for Environmental and Climate Research, Lund University, Sweden*

Abstract

Accurate maps of past land cover and human land use are necessary when studying the impact of anthropogenic land-cover changes on climate. Ideally the maps of past land cover would be separated into naturally occurring vegetation and human induced changes, allowing us to quantify the effect of human land-use on past climate. Here we investigate the possibility of combining regional, fossil pollen based, land-cover reconstructions with, population based, estimates of past human land use. By merging these two datasets and interpolating the pollen based land-cover reconstructions we aim at obtaining maps that provide both past natural land cover and the anthropogenic land-cover changes.

We develop a Bayesian hierarchical model to handle the complex data, using a latent Gaussian Markov random fields (GMRF) for the interpolation. Estimation of the model is based on a block updated Markov chain Monte Carlo (MCMC) algorithm. The sparse precision matrix of the GMRF together with an adaptive Metropolis adjusted Langevin step allows for fast inference. Uncertainties in the land-use predictions are computed from the MCMC posterior samples.

The model uses the pollen based observations to reconstruct three composition of land cover; Coniferous forest, Broadleaved forest and Unforested/Open

land. The unforested land is then further decomposed into natural and human induced openness by inclusion of the estimates of past human land use. The model is applied to five time periods - centred around 1900 CE, 1725 CE, 1425 CE, 1000 and, 4000 BCE over Europe. The results suggest pollen based observations can be used to recover past human land use by adjusting the population based anthropogenic land cover changes estimates.

Key words: Spatial statistics, Gaussian Markov random fields, Dirichlet observations, compositional data, anthropogenic land cover changes, fossil pollen records.

1 Introduction

Human activities mainly influences the climate through the emission of greenhouse gases and anthropogenic land cover changes (ALCC) (Kalnay and Cai, 2003). The effects of both natural and human induced land-cover changes on climate have been investigated in several simulation studies at both global (e.g. Armstrong et al., 2016, Bala et al., 2007, Betts et al., 2007, Brovkin et al., 2002, Christidis et al., 2013, Claussen et al., 2001, Pitman et al., 2009, Pongratz et al., 2009) and regional scales (e.g. Kalnay and Cai, 2003, Strandberg et al., 2014).

Historic ALCC consists mainly of deforestation to allow for agriculture and urbanization (Ruddiman, 2005). For temperate latitudes simulation studies indicate that replacing forests with agricultural land tends to decrease the radiative forcing (and thus temperature) (Bala et al., 2007, Betts et al., 2007), while observational studies show local temperature increases due to urbanization (Kalnay and Cai, 2003). The temperature decreases due to human deforestation are, to some extent, balanced by greenhouse gas emission due to the deforestation (CO_2) and farming practices (Methane) on the deforested land (Kaplan, 2013, Ruddiman, 2005). Earth system models that include dynamic vegetation, allowing for feedback between changes in climate, global CO_2-levels, and vegetation, give an even more complex picture. For these models the effects of ALCC depends on the global CO_2-levels, the climate region, and the natural land cover replaced by human land use (Armstrong et al., 2016).

Comparing historical temperature records with past natural land cover and ALCC might improve our understanding of interactions among climate, land cover, and human land use (Strandberg et al., 2014). However, descriptions of

both past natural land cover (e.g. Brovkin et al., 2002, Hickler et al., 2012, Strandberg et al., 2011) and past ALCC scenarios (e.g. Kaplan et al., 2009, Klein Goldewijk et al., 2011, Pongratz et al., 2009) varies considerably (Gaillard et al., 2010). It was previously shown that fossil pollen records can be used to reconstruct past vegetation and land cover at both local (Sugita, 2007a), regional (Paciorek and McLachlan, 2009, Sugita, 2007b, Sugita et al., 2010), and continental scales (Pirzamanbein et al., 2014).

This paper investigates the possibility of reconstructing both past natural land cover and the ALCC by extending the Bayesian hierarchical model introduced by Pirzamanbein et al. (2015). The fossil pollen data can be used to obtain past land cover (Sugita, 2007b), but does not distinguish naturally open land from deforestation caused by ALCC. Ideally we would like to combine land cover estimates based on fossil pollen records with archaeological data. However, initial studies of available archaeological data revealed a number of potential issues (see discussion in Section 5.1).

To investigate if the modelling is possible we instead used ALCC scenarios (Kaplan et al., 2009, Klein Goldewijk et al., 2011) as an estimate of past ALCC. The resulting model can be seen as an adjustment of the ALCC scenarios based on information in the pollen records (The available data is described in Section 2). The reconstruction is done across Europe for five time periods — centred around 1900, 1725, 1425 CE and 1000, 4000 BCE. These time periods represent important historical periods (recent past, little ice age, black death, late bronze age, and early Neolithic) and are commonly used in both climate modelling and palaeoecological studies. In Section 5.1 we outline one way of extending the model to include archaeological data, and we hope that our results will encourage the development of archaeological databases, that can be used in future modelling.

The model presented here (see Section 3) considers the pollen based land cover data to be Dirichlet observations and the ALCC scenarios to beta observations of underlying latent fields. The spatial structure in the latent fields is modelled using covariates and Gaussian Markov Random Fields (Lindgren et al., 2011). The model is estimated using a Markov chain Monte Carlo (MCMC) algorithm based on the Metropolis Adjusted Langevin algorithm (MALA) (Girolami and Calderhead, 2011). Results are presented in Section 4 and Section 5 concludes the analysis with a discussion.

2 Data

The available data consist of fossil pollen based land cover data, estimates of past human land-use (ALCC scenarios) and potential covariates (elevation and output from a dynamic vegetation model – DVM).

2.1 Pollen based land-cover compositions

Pollen based estimates of three land cover compositions (LCCs), Coniferous forest (C), Broadleaved forest (B) and Unforested land (U), were obtained from the LANDCLIM project (Gaillard et al., 2010) using the REVEALS model (Sugita, 2007b). These three land cover types are commonly used in studies of past climate and climate modelling (Strandberg et al., 2014). REVEALS is a mechanistic model which uses inter-taxonomic differences in pollen productivity, dispersal and the size of sedimentary basins to estimate regional land cover from pollen records. The sedimentary pollen records used by REVEALS are obtained from lakes and bogs and presented as grid based REVEALS estimates for the $1^\circ \times 1^\circ$ grid cells containing sampled lakes and/or bogs (Hellman et al., 2008, showed that the spatial scale of REVEALS reconstructions is around 100×100 km). The resulting land cover data consists of pollen based LCCs for respectively 175, 181, 193, 204 and 196 grid cells during the five time periods centred around 1900, 1725 and 1425 CE, 1000 and 4000 BCE (Trondman et al., 2015). For use in climate modelling these sparse LCC observations can be interpolated to continuous spatial maps (Pirzamanbein et al., 2015). Here we will perform the interpolation while also trying to separate the LCC into natural vegetation and ALCC.

2.2 Anthropogenic land-cover change scenarios

Two anthropogenic land cover change (ALCC) scenarios are used as estimates of human land use: 1) The Kaplan and Krumhardt 2010 scenario (KK10; Kaplan et al., 2009), and 2) The History Database of the Global Environment (HYDE; Klein Goldewijk et al., 2011). KK10 and HYDE are both based on historic human population density estimates, the land needed to feed that population, and soil productivity. To match the pollen records, the two estimates of human land use were upscaled (by averaging) to the $1^\circ \times 1^\circ$ grid cells.

The KK10 and HYDE datasets differ substantially for the older time periods (see Fig. 1), due to differences in assumptions, modelling approaches, and his-

torical records used. In general KK10 gives higher estimates of human land use. Both datasets exhibit substantial local structure.

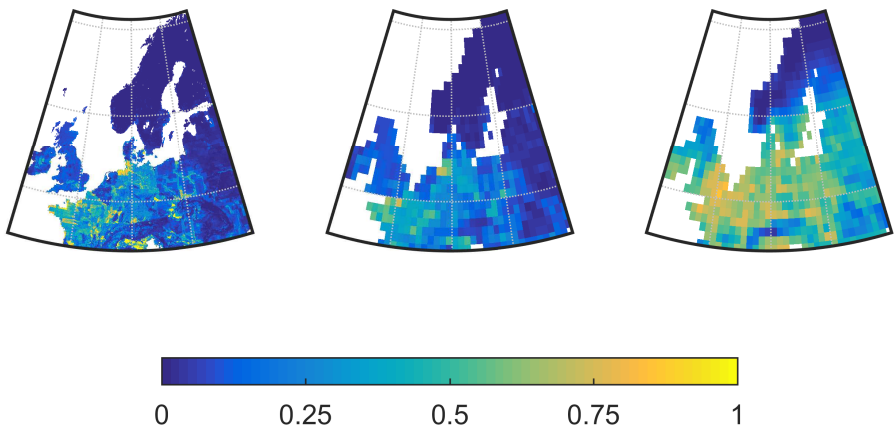


Figure 1: Anthropogenic land cover changes (ALCC) scenarios for 1400 CE. From left to right: The high-resolution (5′ or about 10 km) HYDE ALCC scenario (Klein Goldewijk et al., 2011), HYDE upscaled to 1° resolution matching the pollen data, and the KK10 (Kaplan et al., 2009) ALCC scenario at 1° resolution.

2.3 Covariates

To capture large scale structures in the LCC, covariates consisting of elevation (from the Shuttle Radar Topography Mission[1], Becker et al., 2009) and model based vegetation estimates can be used (Pirzamanbein et al., 2016).

The model based estimates of potential natural vegetation were obtained by running a process-based dynamic vegetation model (DVM), LPJ-GUESS, (Smith et al., 2001) for the study area and specified time periods. LPJ-GUESS estimates the potential natural vegetation based on bio-climatic variables such as temperature, precipitation, and soil types (see Pirzamanbein et al., 2014, for details regarding the LPJ-GUESS runs).

[1] downloaded from `ftp://topex.ucsd.edu/pub/srtm30_plus/` on 2011–09–03

3 Model

For the modelling we assume that each grid cell has a natural LCC, $\mathbf{p}_L = (p_C, p_B, p_U)$, representing the proportion of each grid cell that would be coniferous, broadleaved, or unforested without any human activity. Additionally we let p_H denote the share of each grid cell that is affected by ALCC. Since the ALCC data represents human land use for food production we assume that all human land use can be seen as a replacement of the corresponding proportion of natural land cover with open land. The resulting link between natural and actual land cover, $\mathbf{z} = (z_C, z_B, z_U)$, is

$$\begin{aligned} z_C &= p_C(1 - p_H), \\ z_B &= p_B(1 - p_H), \\ z_U &= p_U(1 - p_H) + p_H, \end{aligned}$$

with the transformation being denoted $\mathbf{z} = h(\mathbf{p}_L, p_H)$ (compare to the covariate adjustments in Pirzamanbein et al., 2014).

The pollen based land cover compositions $\mathbf{L} = (L_C, L_B, L_U)$ are now seen as Dirichlet distributed observations of the actual land cover, $\mathbf{z}$. Similarly the ALCC proportions H are modelled as draws from beta distributions with expectation $p_{H,k}$, where $p_{H,k}$ are perturbations of p_H introduced to handle the (large) differences between the two ALCC datasets (see Figure 1). The resulting model for the pollen and ALCC data given the underlying proportions is

$$\begin{aligned} \mathbf{L}(\mathbf{s})|\alpha, \mathbf{z}(\mathbf{s}) &\sim \mathsf{Dir}\left(\alpha, \mathbf{z}(\mathbf{s})\right), \\ H_k(\mathbf{s})|\lambda, p_{H,k}(\mathbf{s}) &\sim \mathsf{Beta}\left(\lambda p_{H,k}(\mathbf{s}), \lambda(1 - p_{H,k}(\mathbf{s}))\right). \end{aligned} \tag{1}$$

Here $\mathbf{s}$ is the location of each grid cell and α and λ are concentration parameters controlling the uncertainty in the Dirichlet and beta distributions.

We model the grid cell proportions $\mathbf{p}_L(\mathbf{s}) = (p_C(\mathbf{s}), p_B(\mathbf{s}), p_U(\mathbf{s}))$ and $p_H(\mathbf{s})$ as a transformation of an multivariate latent field $\boldsymbol{\eta}(\mathbf{s})$,

$$\mathbf{p}_L(\mathbf{s}) = f\big(\boldsymbol{\eta}_L(\mathbf{s})\big), \qquad\qquad p_H(\mathbf{s}) = g\big(\eta_H(\mathbf{s})\big)$$

with $f : \mathbb{R}^2 \to (0,1)^3$ and $g : \mathbb{R} \to (0,1)$. For f we use the inverse additive

log-ratio transformation (applied for each grid cell, $\mathbf{s}$),

$$\boldsymbol{\eta}_L = \left(\log\left(\frac{p_C}{p_U}\right), \log\left(\frac{p_B}{p_U}\right)\right) = \left(\eta_{L_1}, \eta_{L_2}\right)$$
$$p_\bullet = \begin{cases} \dfrac{\exp(\eta_{L_i})}{1 + \sum_i \exp(\eta_{L_i})} & \text{for } p_C \text{ and } p_B \text{ with } i = 1, 2, \\ \dfrac{1}{1 + \sum_i \exp(\eta_{L_i})} & \text{for } p_U. \end{cases} \tag{2}$$

and for g the inverse logit transformation

$$\eta_H = \log\left(\frac{p_H}{1 - p_H}\right) \qquad \text{and} \qquad p_H = \frac{\exp(\eta_H)}{1 + \exp(\eta_H)}. \tag{3}$$

The components of the latent field $\boldsymbol{\eta}(\mathbf{s})$ are collected into a column vector and modelled using a mean part, $\mathbf{B}\boldsymbol{\beta}$, and a component capturing spatial dependencies $\mathbf{X}$:

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_{L_1} \\ \boldsymbol{\eta}_{L_2} \\ \boldsymbol{\eta}_H \end{bmatrix} = \mathbf{B}\boldsymbol{\beta} + \mathbf{X}.$$

Here $\mathbf{B}$ is a matrix of covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\mathbf{X}$ is a multivariate spatial field.

For $\boldsymbol{\eta}_H$ covariates in $\mathbf{B}$ consist of an intercept and elevation. For $\boldsymbol{\eta}_L$ two possible sets of covariates consisting of either intercept and elevation; or intercept, elevation, and model based vegetation estimates (from LPJ-GUESS) will be evaluated. For the LPJ-GUESS covariates the DVM based 3-compositions of natural potential vegetation were transformed to $\mathbb{R}^2$ using (2), resulting in two covariates, LPJ-GUESS$_{1,2}$. The spatial field, $\mathbf{X}$, is modelled using a Gaussian Markov random field (GMRF, Rue and Held, 2004) with a separable covariance structure,

$$\mathbf{X} \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{\rho} \otimes \mathbf{Q}(\kappa)^{-1}\right)$$

where $\boldsymbol{\rho}$ is a 3×3 covariance matrix, $\mathbf{Q}(\kappa)$ is the precision matrix of a GMRF that approximates fields with Matérn covariance function (Lindgren and Rue, 2013, Lindgren et al., 2011), and κ governs the range of the spatial dependence.

To handle the differences between the KK10 and HYDE data, perturbed proportions of human land use $p_{H,k}(\mathbf{s})$ were introduced in the data model, (1). These

perturbations are created by adding random effects to the $\boldsymbol{\eta}_H$-field; $p_{H,k}(\mathbf{s})$ is computed from $\eta_{H,k}(\mathbf{s}) = \eta_H(\mathbf{s}) + \varepsilon_k$ using (3) where $\varepsilon_k \sim \mathsf{N}\left(0, \tau_\varepsilon^{-1}\right)$. Note that ε_k are common terms added to the entire field, an attempt to use different random effects for each grid cell, i.e. $\varepsilon_k(\mathbf{s})$, resulted in an unidentifiable model.

The full hierarchical model is illustrated in Figure 2. The final part of the model is to specify suitable priors, following (Pirzamanbein et al., 2015) we use wide priors for α and λ; conjugate priors for $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$; and for κ we pick a prior appropriate to the size of our spatial domain (Fuglstad et al., 2016). Finally we pick a conjugate prior for τ_ε since this, similar to $\boldsymbol{\rho}$, allows for simple MCMC updates. The resulting priors are

$$\begin{aligned}
&\alpha \sim \Gamma(1.5, 0.1), && \lambda \sim \Gamma(1.5, 0.1),\\
&\boldsymbol{\beta} \sim \mathsf{N}\left(0, \mathbb{I} \cdot 10^{-3}\right), && \tau_\varepsilon \sim \Gamma(1.5, 0.1),\\
&\kappa \sim \Gamma\left(1, \frac{\log(100)}{\sqrt{8}}\right), && \boldsymbol{\rho} \sim \mathsf{IW}(\mathbb{I}, 10).
\end{aligned}$$

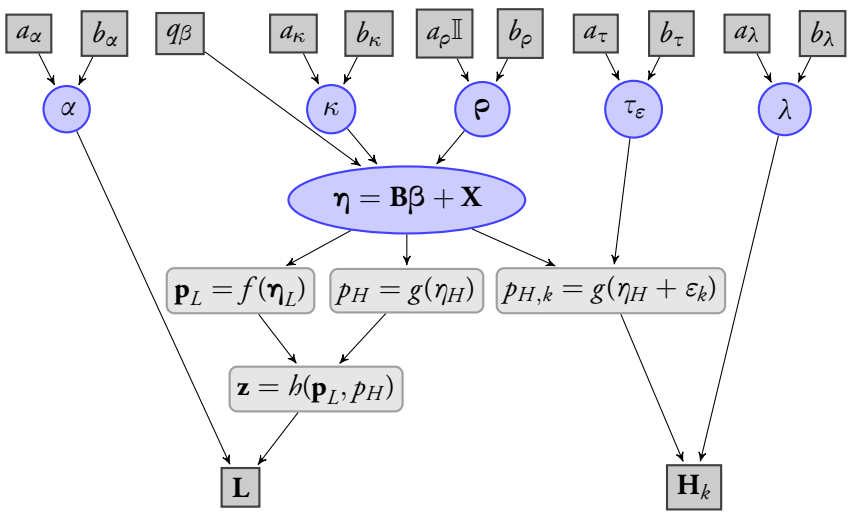


Figure 2: Directed acyclic graph describing the conditional dependencies in the hierarchical model.

3.1 Estimation using MCMC

To estimate model parameters and reconstruct the latent field we use a block-updated MCMC algorithm. In the first block the latent fields – $\boldsymbol{\eta}$, $\boldsymbol{\beta}$, and ε_k – and the Dirichlet and beta concentration parameters – α and λ – are updated using a MALA proposal (Girolami and Calderhead, 2011). In the second block, we update the range parameter of the GMRF – κ – using a random walk in log scale and the covariance matrix – $\boldsymbol{\rho}$ – using the conjugacy (conditioned on κ). Finally τ_ε is updated using the conjugate posterior. In each iteration the MCMC alternates between these three blocks. To get the desired acceptance rate we use an adaptive scheme (Andrieu and Thoms, 2008) where the step size of the MALA proposal and the random walk are adjusted to maintain 57% and 40% acceptance rate, respectively (Roberts et al., 2001). This MCMC is an extension of the implementation, for a simpler model, described by Pirzamanbein et al. (2015).

We ran 100 000 MCMC iterations with a burn-in sample size of 10 000 to estimates the parameters of each model. The MCMC chain plots show convergence and good mixing of the parameters.

4 Results and discussion

The reconstruction of human land use, potential natural vegetation and land cover compositions are shown in Figure 3 for the 1425 CE time period. The results for the other time periods are available in Appendix B. In general, the reconstructions capture the variability in the observed datasets. The human land use reconstructions mostly capture the spatial patterns of KK10 while the amount of land use is closer to HYDE. Moreover, the model with only elevation as covariates estimates slightly higher amounts of human land use compared to the model also including LPJ-GUESS as covariates.

The estimates of ε_k for HYDE and KK10 (Figure 4) also indicate that the human land use reconstructions are, on average, closer to HYDE than KK10 for all the time periods. The difference between HYDE and KK10, as captured by ε_k, increases for older time periods (see Figure 11 in Appendix. C). The estimates of ε_k are higher when the model includes both elevation and LPJ-GUESS as covariates compared to the model only including elevation. This is in accordance with the higher estimates of human land use in the model containing only elevation.

The uncertainties in the human land use reconstructions denote higher vari-

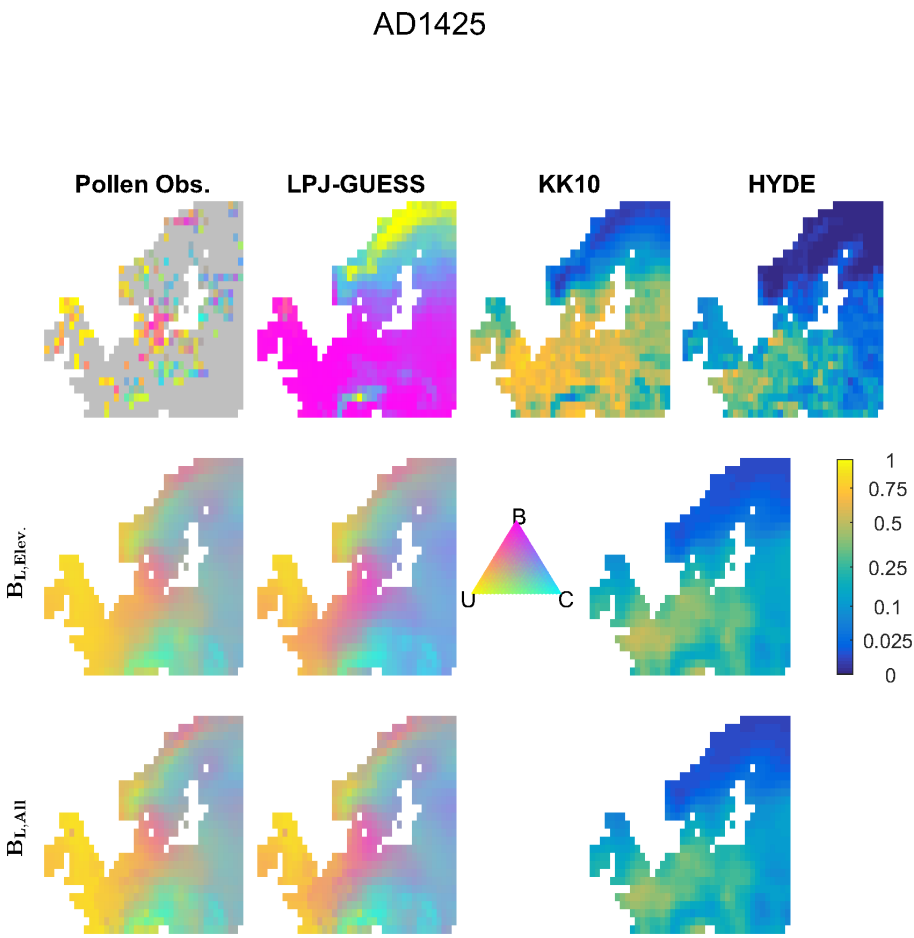


Figure 3: The observation datasets (row 1) and the reconstructions using two different sets of covariates (row 2 and 3) for 1425 CE. From left to right: land cover composition, natural land cover, and human land use.

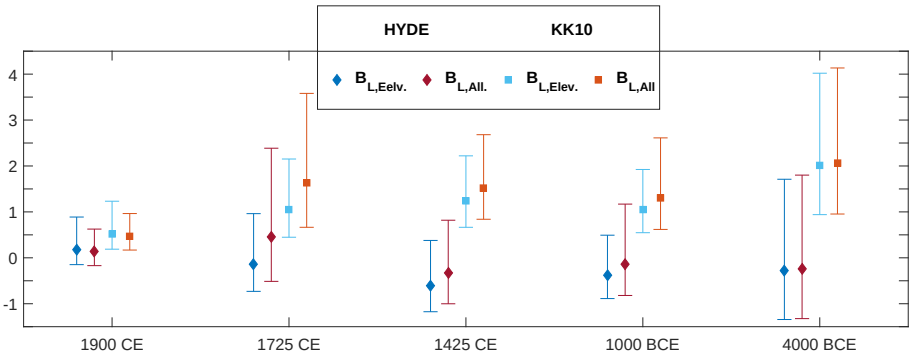


Figure 4: Estimated ε_k for HYDE and KK10 and corresponding 95% confidence intervals for all time periods. The blue color represents model includes only elevation and red color represents model include both elevation and LPJ-GUESS.

ation in the model with LPJ-GUESS as covariates than the model with only elevation (Figure 11 in Appendix. C). The uncertainty in the compositional reconstructions, i.e. natural potential land cover and land cover composition, are computed using transformed elliptical confidence regions (Pirzamanbein et al., 2015). The results together with confidence intervals for human land use are illustrated in Figure 5 for three locations during the 1425 CE time period. The confidence regions are based on the model using only elevation, in order to allow a comparison between the NLC estimates and LPJ-GUESS. The selected point in the Baltic (column 1 in Figure 5) represents a location with contrasting values in the different data sources, i.e. about 70% of coniferous forest in LCC, 70% of broadleaved forest in LPJ-GUESS, and 40% or 10% of human land use in KK10 and HYDE respectively. The differences among the data sources are balanced in the reconstruction of LCC, NLC and human land use. In contrast, when the differences are smaller the confidence regions include the observations quiet well (columns 2 and 3 in Figure 5). The selected point in Scotland (column 3 in Figure 5) shows the improvement of the NLC reconstruction compared to the LPJ-GUESS estimate. The reconstruction suggests that the 80% of unforested land consist of 10% human land use while LPJ-GUESS suggests 30% unforested land and 70% boardleaved forest.

A leave out validation is used to evaluate the performance of the model. The validation is performed by randomly removing 10% of observed grid cells in the

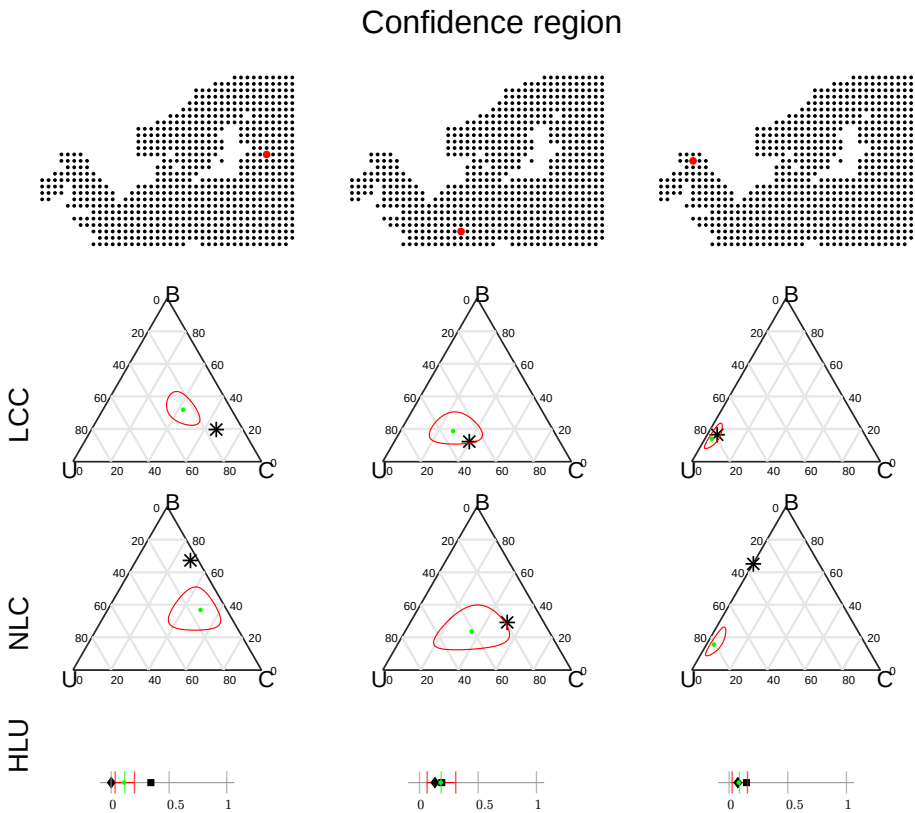


Figure 5: The reconstruction and prediction regions for three locations for land cover composition (LCC), natural Land cover (NLC) and human land-use (HLU) for 1425 CE. For LCC and NLC (rows 2 and 3) the observations, pollen based REVEALS reconstructions and LPJ-GUESS output respectively, are marked with (∗). For HLU (row 4) the two ALCC observations are given by HYDE (⧫) and KK10 (■). For all figures the green dots indicate estimated values and the red lines represent the corresponding confidence regions. All estimates and confidence regions are based on the model without LPJ-GUESS$_{1,2}$ as covariates.

LCC and ALCC data and reconstruct these values based on the remaining observations. The resulting land cover reconstructions are compared to LCC using average compositional distance (ACD; see Aitchison et al., 2000, Pirzamanbein et al., 2015), and the human land-use reconstructions are compared to both KK10 and HYDE using root mean squared error (RMSE). Comparing the ACD and RMSE (Table 1), there is no general preference in for any of the two models with different covariates. As has previously been noted the HLU estimates are, in general, closer to HYDE than to KK10.

	ACD		**RMSE**			
	REV		KK10		HYDE	
	$\mathbf{B}_{\text{All}}$	$\mathbf{B}_{\text{Elev.}}$	$\mathbf{B}_{\text{All}}$	$\mathbf{B}_{\text{Elev.}}$	$\mathbf{B}_{\text{All}}$	$\mathbf{B}_{\text{Elev.}}$
1900 CE	1.01	**0.78**	0.13	**0.09**	0.14	0.14
1725 CE	1.26	**1.15**	0.16	0.20	0.15	**0.12**
1425 CE	**1.40**	**1.40**	0.17	0.20	**0.15**	0.17
1000 BCE	**1.02**	1.16	0.10	0.13	**0.07**	**0.07**
4000 BCE	1.34	**0.99**	0.15	0.12	**0.05**	**0.05**

Table 1: Leave out validation results for models with two different sets of covariates, $\mathbf{B}_{\text{All}}$,$\mathbf{B}_{\text{Elev.}}$ and all time periods. The reconstructions of land cover compositions (LCC) are compared using average compositional distances (ACD). The human land use (HLU) reconstructions are compared using root mean square error (RMSE). The bold number indicates the lowest value in the row for LCC and HLU.

5 Conclusion

In this paper, we developed a Bayesian hierarchical model to reconstruct the past human land-use for five time periods centred around 1900 CE, 1725 CE, 1425 CE, 1000 BCE and 4000 BCE. The reconstructions are based on combination of pollen based land cover compositions (Trondman et al., 2015) and population based anthropogenic land cover changes (ALCC) estimates.

Due to discrepancies between the past ALCC estimates, the model uses two different datasets of human land use: anthropogenic land cover changes scenario of (; KK10 Kaplan et al., 2009) and historic data base of global environment (HYDE; Klein Goldewijk et al., 2011). The past human land use reconstruction capture the spatial patterns of KK10 while being closer in value to the proportions of HYDE. This suggests that pollen based LCC can be used to adjust the existing population based human land use to match observed past vegetation patterns and recover past human land use from pollen based LCC.

We note that the model would allow the inclusion of additional anthropogenic land cover changes scenarios and it would be interesting to also include archaeological data. However, our initial attempts to use archaeological data have so far, as described below, been unsuccessful.

5.1 Including archaeological data in the model

We initially considered using archaeological data, instead of the ALCC scenarios, as a measure of human land use. Given an archaeological dataset containing the locations of relevant archaeological finds during each of the five time periods we would replace the β-observations of the ALCC scenarios with a point process (Simpson et al., 2016) over the archaeological finds. The base idea being that more finds, in a given region, would correspond to a higher human activity and thus a higher proportion of ALCC.

One possible model would be an exponential link-function between the latent field, $\boldsymbol{\eta}_H$, and the intensity, $\boldsymbol{\lambda}$, of the point process for the archaeological finds, e.g.

$$\begin{aligned} \boldsymbol{\lambda} &= \exp\left(\boldsymbol{\eta}_H\right) \\ \log \mathsf{P}(\mathbf{A}|\lambda) &= |\Omega| - \int_\Omega \lambda(s)\,\mathrm{d}s + \sum_{i=1}^{n} \log \lambda(s_i) \end{aligned}$$

where $\mathbf{A} = \{s_i\}$ are the locations of the archaeological finds. Since the point process provides the relative frequency of events, the latent field, $\boldsymbol{\eta}_H$, might only be determined up to an additive constant. To make the model identifiable the ALCC scenarios could still be needed, either as observations or as covariates. While it would be very interesting to investigate this model we have been unable to find a suitable archaeological dataset.

For us, a large detrimental factor to the use of archaeological data has been our inability to find archaeological databases covering the entire study area. One option considered was to restrict the modelling to Sweden using the *Fornsök*-database[2] maintained by the Swedish National Heritage Board. This database contains information regarding roughly 1.7 million finds, but is incomplete with data contributions largely depending on the local municipalities (*kommuner*).

An initial search of the database resulted in 68 000 dated finds marked as relating to agricultural and/or settlement activities. And an additional 54 000 finds in these categories without any dating information. The spatial information regarding finds is good ($\pm$250 m, i.e. much smaller than the spatial resolution of the pollen based LCCs). However, the dating information ranges from very good (based on C_{14} or dendrochronology) to rather inexact. With most of the finds being dated based on typology, i.e. as belonging to one (or several) of 5 time periods. The wide ranges of possible dates and the uncertainty regarding selection bias due to differing priorities among the contributing municipalities makes the data unsuitable for our purposes (see Fig. 6).

Acknowledgement

The research presented in this paper is a contribution to the two Swedish strategic research areas Biodiversity and Ecosystems in a Changing Climate (BECC), and ModElling the Regional and Global Earth system (MERGE).

We thank M.-J. Gaillard and A. Poska for providing the pollen based land cover data complied by LAND Cover-CLIMate interactions in NW Europe during the Holocene (LANDCLIM) project, natural vegetation cover form LPJ-GUESS, and anthropogenic land cover changes of KK10 data bases.

[2] http://www.raa.se/in-english/about-fornsok/

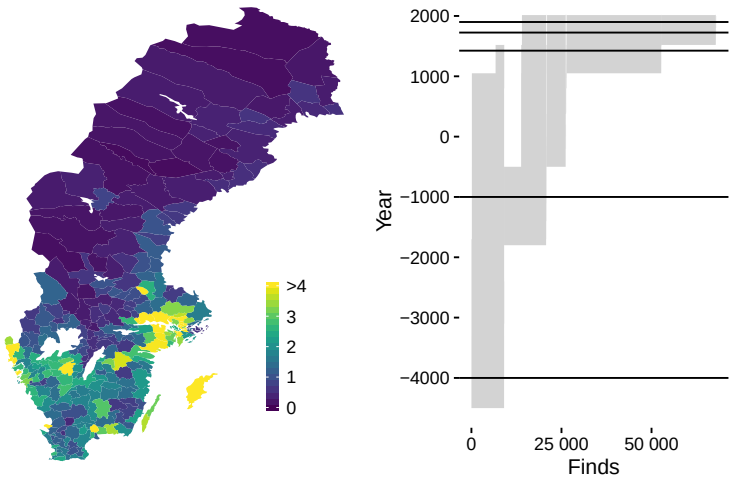


Figure 6: Overview of the Swedish archaeological data. The left pane shows the total number of finds per square kilometer for each of the 290 municipalities (*kommuner*) of Sweden. In the right pane the grey area indicates the dating range given for each archaeological find. The five time periods for which we have pollen data are indicated by the horizontal black lines.

A Computation for MALA proposal

For MALA proposal, the computation of the log density, first derivatives and expected Fisher information of the Beta distribution are required. The Fisher information is the negative expectation with respect to observations of the second and partial derivatives of the log density with respect to parameters and latent field.

A.1 Beta distribution computations

The Beta density is

$$\mathsf{P}(\mathbf{y}|\lambda,\mathbf{p}) = \frac{\Gamma(\lambda)}{\Gamma(\lambda\mathbf{p})\Gamma(\lambda(1-\mathbf{p}))}\mathbf{y}^{\lambda\mathbf{p}-1}(1-\mathbf{y})^{\lambda(1-\mathbf{p})-1} \quad \lambda>0,\ \mathbf{p}\in(0,1),$$

therefore the log density becomes

$$\begin{aligned} l = \log \mathsf{P}(\mathbf{y}|\lambda,\mathbf{p}) = &\log\Gamma(\lambda) - \log\Gamma(\lambda\mathbf{p}) - \log\Gamma(\lambda(1-\mathbf{p})) \\ &+ (\lambda\mathbf{p}-1)\log\mathbf{y} + (\lambda(1-\mathbf{p})-1)\log(1-\mathbf{y}). \end{aligned}$$

The first derivatives with respect to the parameters, λ and $\mathbf{p}$ are

$$\begin{aligned} \frac{\partial l}{\partial\lambda} &= \psi(\lambda) - \mathbf{p}\psi(\lambda\mathbf{p}) - (1-\mathbf{p})\psi(\lambda(1-\mathbf{p})) + \mathbf{p}\log\mathbf{y} + (1-\mathbf{p})\log(1-\mathbf{y}), \\ \frac{\partial l}{\partial\mathbf{p}} &= -\lambda\psi(\lambda\mathbf{p}) + \lambda\psi(\lambda(1-\mathbf{p})) + \lambda\log\mathbf{y} - \lambda\log(1-\mathbf{y}). \end{aligned}$$

The second and partial derivatives are

$$\begin{aligned} \frac{\partial^2 l}{\partial\lambda^2} =& \psi'(\lambda) - \mathbf{p}^2\psi'(\lambda\mathbf{p}) - (1-\mathbf{p})^2\psi'(\lambda(1-\mathbf{p})), \\ \frac{\partial^2 l}{\partial\mathbf{p}^2} =& -\lambda^2\psi'(\lambda\mathbf{p}) - \lambda^2\psi'(\lambda(1-\mathbf{p})), \\ \frac{\partial^2 l}{\partial\mathbf{p}\partial\lambda} =& -\psi(\lambda\mathbf{p}) - \lambda\mathbf{p}\psi'(\lambda\mathbf{p}) + \psi(\lambda(1-\mathbf{p})) + \lambda(1-\mathbf{p})\psi'(\lambda(1-\mathbf{p})) \\ &+ \log\mathbf{y} - \log(1-\mathbf{y}). \end{aligned}$$

The symmetric Fisher information is

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{\lambda,\lambda} & \mathcal{I}_{\lambda,\mathbf{p}} \\ & \mathcal{I}_{\mathbf{p},\mathbf{p}} \end{bmatrix} = -\mathsf{E}_{\mathbf{y}} \begin{bmatrix} \dfrac{\partial^2 l}{\partial \lambda^2} & \dfrac{\partial^2 l}{\partial \mathbf{p} \partial \lambda} \\ & \dfrac{\partial^2 l}{\partial \mathbf{p}^2} \end{bmatrix}$$

with elements

$$\begin{aligned} \mathcal{I}_{\lambda,\lambda} =& -\psi'(\lambda) + \mathbf{p}^2\psi'(\lambda\mathbf{p}) + (1-\mathbf{p})^2\psi'(\lambda(1-\mathbf{p})) \\ \mathcal{I}_{\mathbf{p},\mathbf{p}} =& \lambda^2\psi'(\lambda\mathbf{p}) + \lambda^2\psi'(\lambda(1-\mathbf{p})) \\ \mathcal{I}_{\lambda,\mathbf{p}} =& -\psi(\lambda\mathbf{p}) - \lambda\mathbf{p}\psi'(\lambda\mathbf{p}) + \psi(\lambda(1-\mathbf{p})) + \lambda(1-\mathbf{p})\psi'(\lambda(1-\mathbf{p})) \\ & + \log\mathbf{y} - \log(1-\mathbf{y}). \end{aligned}$$

Since $\mathsf{E}(\log\mathbf{y}) = \psi(\lambda\mathbf{p}) - \psi(\lambda)$, $\mathcal{I}_{\lambda,\mathbf{p}}$ simplifies to

$$\mathcal{I}_{\lambda,\mathbf{p}} = \lambda\mathbf{p}\psi'(\lambda\mathbf{p}) - \lambda(1-\mathbf{p})\psi'(\lambda(1-\mathbf{p})).$$

B Maps of reconstructed land cover and human land use

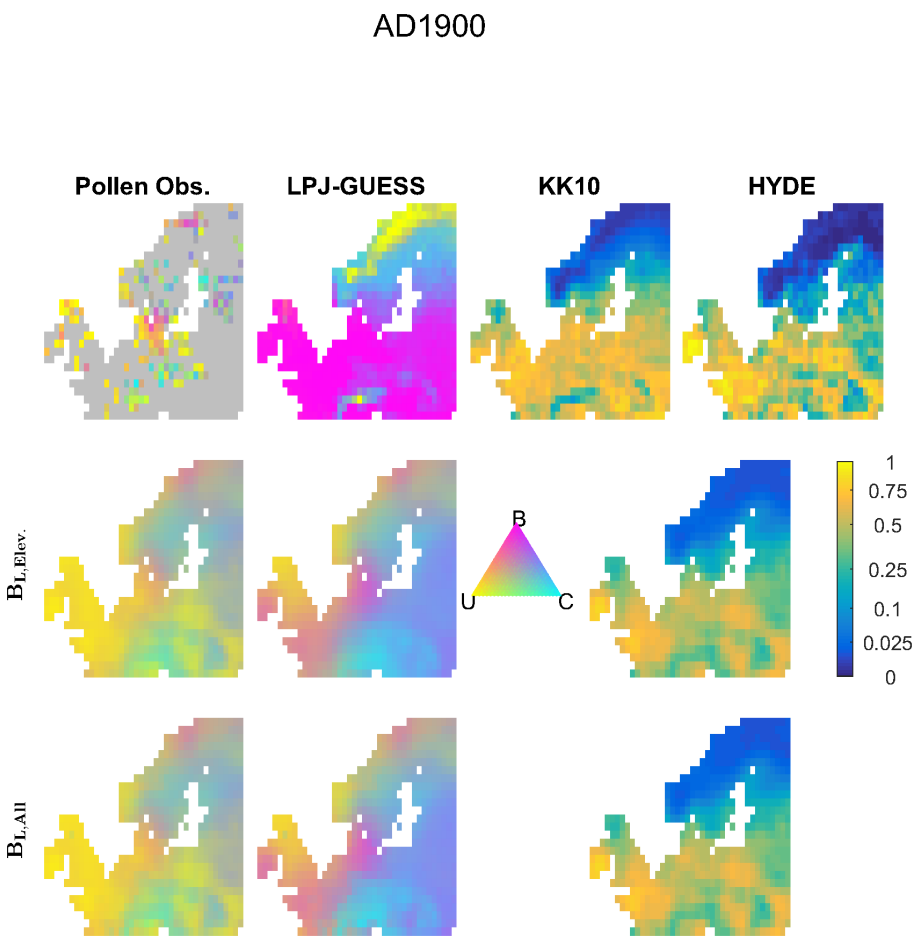


Figure 7: The observation datasets (row 1) and the reconstructions using two different sets of covariates (row 2 and 3) for 1900 CE. From left to right: land cover composition, natural land cover, and human land use.

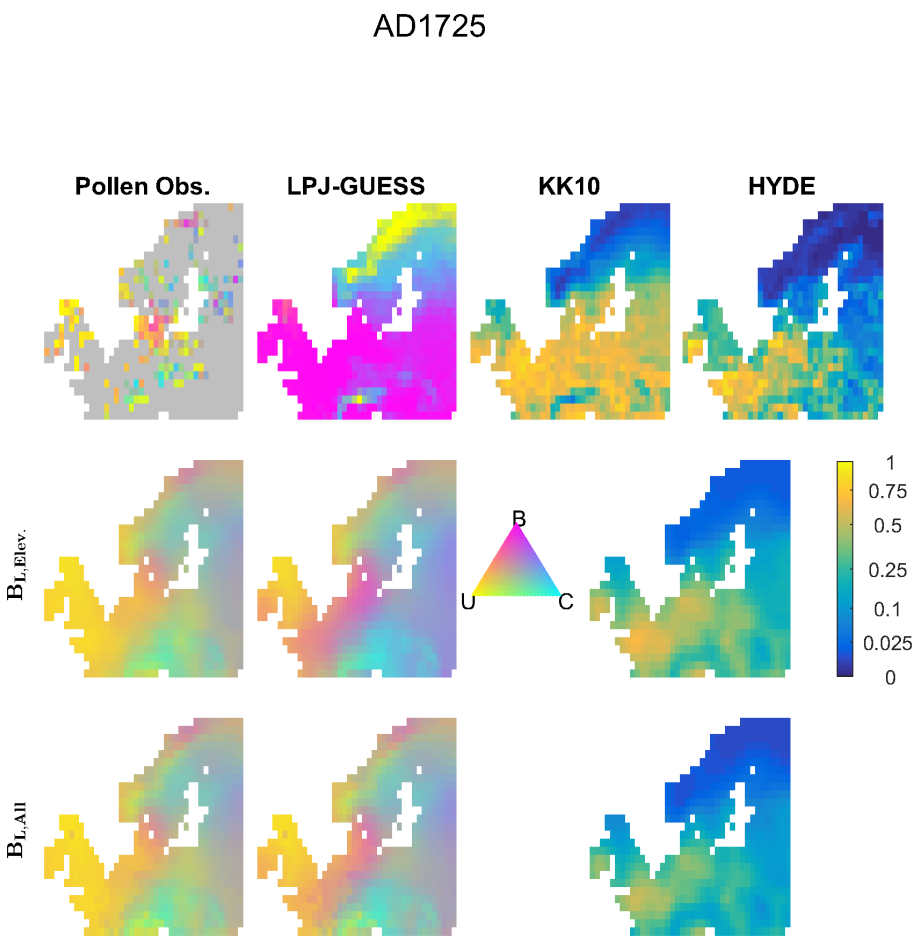


Figure 8: The observation datasets (row 1) and the reconstructions using two different sets of covariates (row 2 and 3) for 1725 CE. From left to right: land cover composition, natural land cover, and human land use.

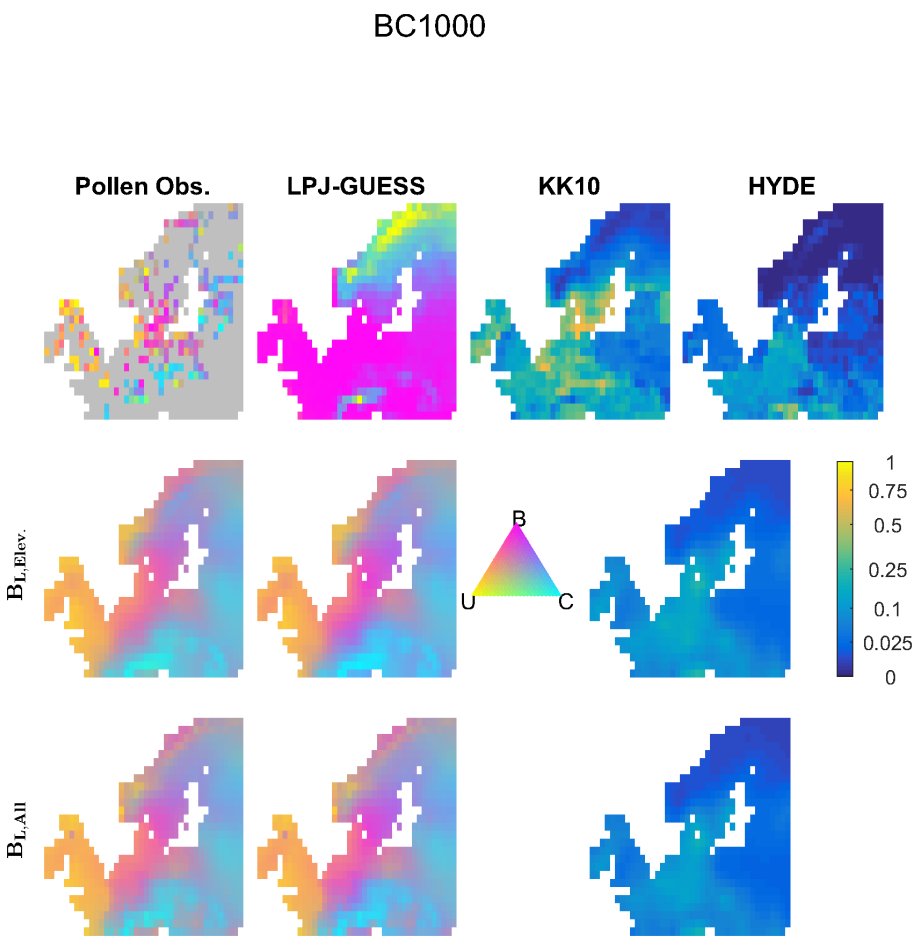


Figure 9: The observation datasets (row 1) and the reconstructions using two different sets of covariates (row 2 and 3) for 1000 BCE. From left to right: land cover composition, natural land cover, and human land use.

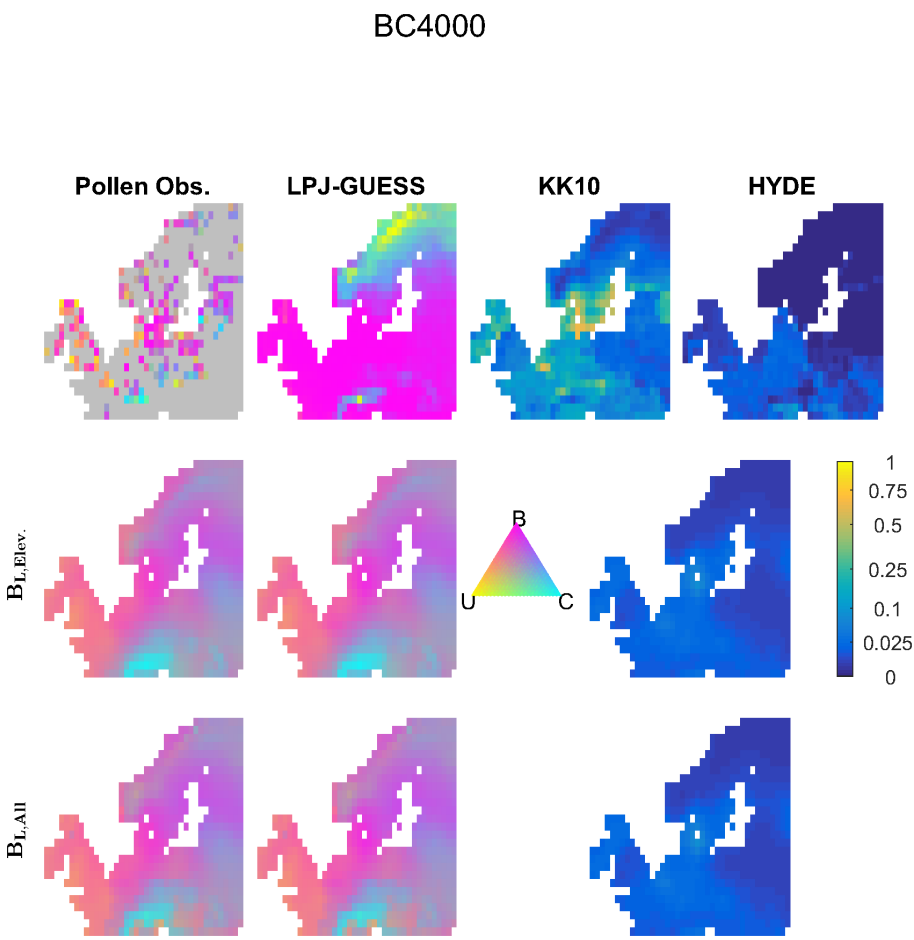


Figure 10: The observation datasets (row 1) and the reconstructions using two different sets of covariates (row 2 and 3) for 4000 BCE. From left to right: land cover composition, natural land cover, and human land use.

C Uncertainties in land use reconstruction

The desription of the figure in this appendix is as follows,
95% confidence interval for human land use reconstructions for all time periods. From left to right: HYDE observations, KK10 observations, lower bound and upper bound for reconstruction of the model with only elevation as covariates, $\mathbf{B}_{\text{Elev.}}$, and lower bound and upper bound for reconstructions of the model with both elevation and LPJ-GUESS as covariates, $\mathbf{B}_{\text{All}}$.

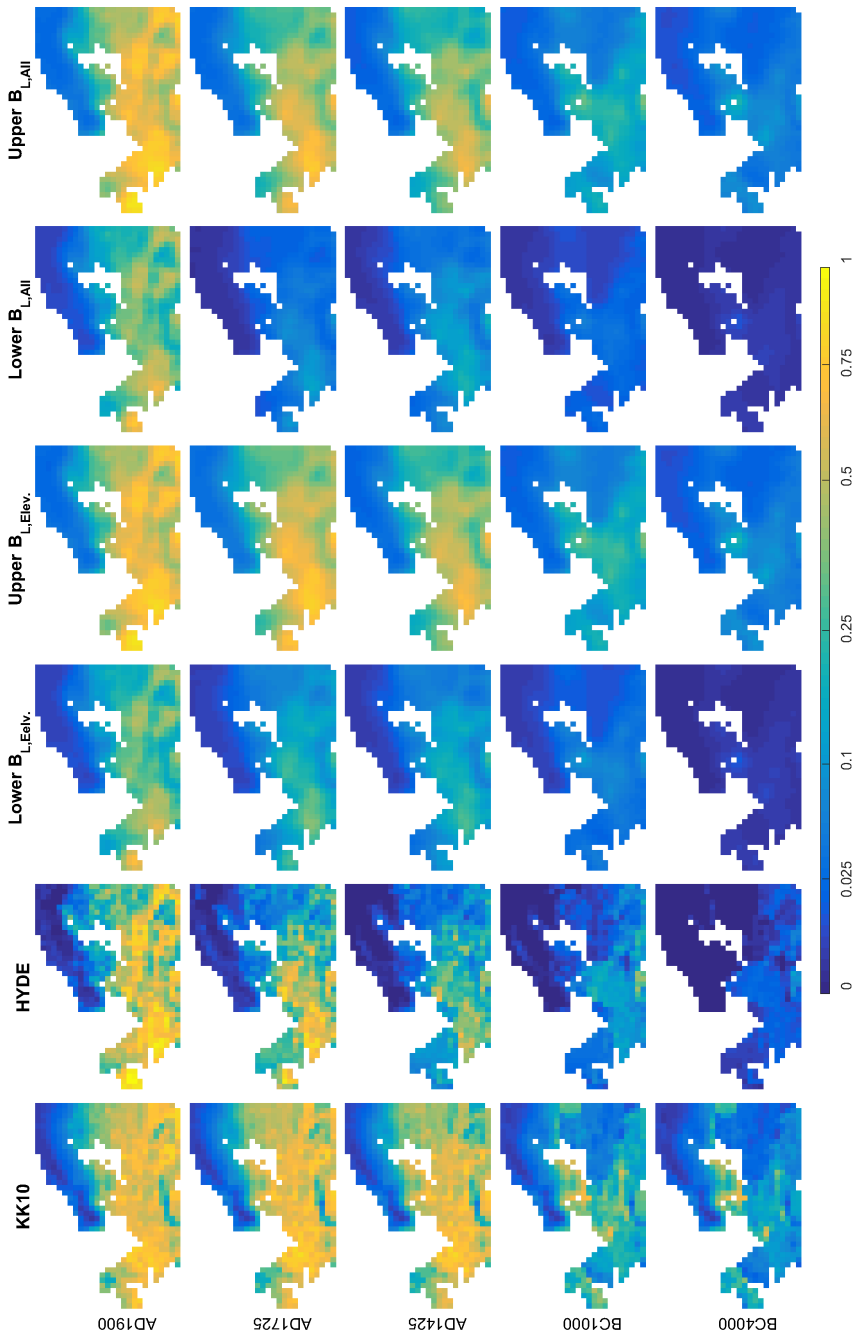


Figure 11: HYDE and KK10 observations and confidence bound for humanland use reconstructions, see page 176.

References

J. Aitchison, C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn. Logratio analysis and compositional distance. *Math. Geol.*, 32(3):271–275, 2000.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. and Comput.*, 18(4):343–373, 2008.

E. Armstrong, P. Valdes, J. House, and J. Singarayer. The role of CO2 and dynamic vegetation on the impact of temperate land-use change in the HadCM3 coupled climate model. *Earth Interactions*, 20(10):1–20, 2016.

G. Bala, K. Caldeira, M. Wickett, T. J. Phillips, D. B. Lobell, C. Delire, and A. Mirin. Combined climate and carbon-cycle effects of large-scale deforestation. 104(16):6550–6555, 2007.

J. J. Becker, D. T. Sandwell, W. H. F. Smith, J. Braud, B. Binder, J. Depner, D. Fabre, J. Factor, S. Ingalls, S. H. Kim, R. Ladner, K. Marks, S. Nelson, A. Pharaoh, G. Sharman, R. Trimmer, J. VonRosenburg, G. Wallace, and P. Weatherall. Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Mar. Geod.*, 32(4):355–371, 2009.

R. A. Betts, P. D. Falloon, K. K. Goldewijk, and N. Ramankutty. Biogeophysical effects of land use on climate: Model simulations of radiative forcing and large-scale temperature change. *Agricultural and Forest Meteorology*, 142(2–4):216–233, 2007.

V. Brovkin, J. Bendtsen, M. Claussen, A. Ganopolski, C. Kubatzki, V. Petoukhov, and A. Andreev. Carbon cycle, vegetation, and climate dynamics in the Holocene: Experiments with the CLIMBER-2 model. *Global. Biogeochem. Cy.*, 16 (4):1139, 2002.

N. Christidis, P. A. Stott, G. C. Hegerl, and R. A. Betts. The role of land use change in the recent warming of daily extreme temperatures. *Geophys. Res. Lett.*, 40(3):589–594, 2013.

M. Claussen, V. Brovkin, and A. Ganopolski. Biogeophysical versus biogeochemical feedbacks of large-scale land cover change. *Geophys. Res. Lett.*, 28(6):1011–1014, 2001.

G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Interpretable priors for hyperparameters for Gaussian Random Fields. Technical Report 1503.00256v2, arXiv, 2016. URL `http://arxiv.org/abs/1503.00256v2`.

M.-J. Gaillard, S. Sugita, F. Mazier, A.-K. Trondman, A. Brostrom, T. Hickler, J. O. Kaplan, E. Kjellström, U. Kokfelt, P. Kuneš, , C. Lemmen, P. Miller, J. Olofsson, A. Poska, M. Rundgren, B. Smith, G. Strandberg, R. Fyfe, A. Nielsen, T. Alenius, L. Balakauskas, L. Barnekov, H. Birks, A. Bjune, L. Björkman, T. Giesecke, K. Hjelle, L. Kalnina, M. Kangur, W. van der Knaap, T. Koff, P. Lagerås, M. Latałowa, M. Leydet, J. Lechterbeck, M. Lindbladh, B. Odgaard, S. Peglar, U. Segerström, H. von Stedingk, and H. Seppä. Holocene land-cover reconstructions for studies on land cover-climate feedbacks. *Clim. Past.*, 6:483–499, 2010.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 73(2):123–214, 2011.

S. E. Hellman, M.-j. Gaillard, A. Broström, and S. Sugita. Effects of the sampling design and selection of parameter values on pollen-based quantitative reconstructions of regional vegetation: a case study in southern Sweden using the REVEALS model. *Veg. Hist. Archaeobot.*, 17(5):445–459, 2008.

T. Hickler, K. Vohland, J. Feehan, P. A. Miller, B. Smith, L. Costa, T. Giesecke, S. Fronzek, T. R. Carter, W. Cramer, I. Kühn, and M. T. Sykes. Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global. Ecol. Biogeogr.*, 21(1):50–63, 2012.

E. Kalnay and M. Cai. Impact of urbanization and land-use change on climate. *Nature*, 423(6939):528–531, 2003.

J. O. Kaplan. From forest to farmland and meadow to metropolis: What role for humans in explaining the enigma of Holocene CO2 and methane concentrations? In *EGU General Assembly Conference Abstracts*, volume 15, page 886, 2013.

J. O. Kaplan, K. M. Krumhardt, and N. Zimmermann. The prehistoric and preindustrial deforestation of Europe. *Quaternary. Sci. Rev.*, 28(27):3016–3034, 2009.

K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Global. Ecol. Biogeogr.*, 20(1):73–86, 2011.

F. Lindgren and H. Rue. Bayesian spatial and spatio-temporal modelling with R-INLA. Submitted to J. Stat. Softw., 2013. URL `http://www.math.ntnu.no/inla/r-inla.org/papers/jss/lindgren.pdf`.

F. Lindgren, R. Håvard, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc. Ser. B*, 73(4):423–498, 2011.

C. J. Paciorek and J. S. McLachlan. Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Am. Statist. Assoc.*, 104(486):608–622, 2009.

B. Pirzamanbein, J. Lindström, A. Poska, S. Sugita, A.-K. Trondman, R. Fyfe, F. Mazier, A. B. Nielsen, J. O. Kaplan, A. E. Bjune, H. J. B. Birks, T. Giesecke, M. Kangur, M. Latałowa, L. Marquer, B. Smith, and M.-J. Gaillard. Creating spatially continuous maps of past land cover from point estimates: A new statistical approach applied to pollen data. *Ecol. Complex.*, 20(0):127 – 141, 2014.

B. Pirzamanbein, J. Lindström, A. Poska, and M.-J. Gaillard. Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *arXiv preprint arXiv:1511.06417*, 2015.

B. Pirzamanbein, A. Poska, and J. Lindström. Analysing the sensitivity of pollen based land cover maps to different auxiliary variables. in preparation, 2016.

A. Pitman, N. de Noblet-Ducoudré, F. Cruz, E. Davin, G. Bonan, V. Brovkin, M. Claussen, C. Delire, L. Ganzeveld, V. Gayler, B. J. J. M. van den Hurk, P. J. Lawrence, M. K. van der Molen, C. Müller, C. H. Reick, S. I. Seneviratne, B. J. Strengers, and A. Voldoire. Uncertainties in climate responses to past land cover change: First results from the LUCID intercomparison study. *Geophys. Res. Lett.*, 36(14), 2009.

J. Pongratz, C. Reick, T. Raddatz, and M. Claussen. Effects of anthropogenic land cover change on the carbon cycle of the last millennium. *Global. Biogeochem. Cy.*, 23(4):GB4001, 2009.

G. O. Roberts, J. S. Rosenthal, et al. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

W. F. Ruddiman. How did humans first alter global climate? *Sci. Am.*, March 2005:34–41, 2005.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2004.

D. Simpson, J. Illian, F. Lindgren, S. Sorbye, and H. Rue. Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103 (1):49–70, 2016.

B. Smith, I. C. Prentice, and M. T. Sykes. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Global. Ecol. Biogeogr.*, 10(6):621–637, 2001.

G. Strandberg, J. Brandefelt, E. Kjellström, and B. Smith. High-resolution regional simulation of last glacial maximum climate in Europe. *Tellus. A*, 63(1): 107–125, 2011.

G. Strandberg, E. Kjellström, A. Poska, S. Wagner, M.-J. Gaillard, A.-K. Trondman, A. Mauri, B. A. S. Davis, J. O. Kaplan, H. J. B. Birks, A. E. Bjune, R. Fyfe, T. Giesecke, L. Kalnina, M. Kangur, W. O. van der Knaap, U. Kokfelt, P. Kuneš, M. Latał owa, L. Marquer, F. Mazier, A. B. Nielsen, B. Smith, H. Seppä, and S. Sugita. Regional climate model simulations for Europe at 6 and 0.2 k bp: sensitivity to changes in anthropogenic deforestation. *Clim. Past.*, 10(2):661–680, 2014.

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

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

S. Sugita, T. Parshall, R. Calcote, and K. Walker. Testing the landscape reconstruction algorithm for spatially explicit reconstruction of vegetation in northern Michigan and Wisconsin. *Quaternary. Res.*, 74(2):289–300, 2010.

A.-K. Trondman, M.-J. Gaillard, F. Mazier, S. Sugita, R. Fyfe, A. B. Nielsen, C. Twiddle, P. Barratt, H. J. B. Birks, A. E. Bjune, L. Björkman, A. Broström, C. Caseldine, R. David, J. Dodson, W. Dörfler, E. Fischer, B. van Geel, T. Giesecke, T. Hultberg, L. Kalnina, M. Kangur, P. van der Knaap, T. Koff, P. Kuneš, P. Lagerås, M. Latałowa, J. Lechterbeck, C. Leroyer, M. Leydet, M. Lindbladh, L. Marquer, F. J. G. Mitchell, B. V. Odgaard, S. M. Peglar, T. Persson, A. Poska, M. Rösch, H. Seppä, S. Veski, and L. Wick. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. *Glob. Change Biol.*, 21(2):676–697, 2015.

Doctoral Theses Published in Environmental Science Lund University

Georg K.S. Andersson (2012) Effects of farming practice on pollination across space and time. Department of Biology/Center for Environmental and Climate Research

Anja M. Ödman (2012) Disturbance regimes in dry sandy grasslands – past, present and future. Department of Biology/Center for Environmental and Climate Research

Johan Genberg (2013) Source apportionment of carbonaceous aerosol. Department of Physics/Center for Environmental and Climate Research

Petra Bragée (2013) A palaeolimnological study of the anthropogenic impact on dissolved organic carbon in South Swedish lakes. Department of Geology/Center for Environmental and Climate Research

Estelle Larsson (2013) Sorption and transformation of anti-inflammatory drugs during wastewater treatment. Department of Chemistry/Center for Environmental and Climate Research

Magnus Ellström (2014) Effects of nitrogen deposition on the growth, metabolism and activity of ectomycorrhizal fungi. Department of Biology/Center for Environmental and Climate Research

Therese Irminger Street (2015) Small biotopes in agricultural landscapes: importance for vascular plants and effects on management. Department of physical geography and ecosystem science/ Department of Biology/Center for Environmental and Climate Research

Helena I. Hanson (2015) Natural enemies: Functional aspects of local management in agricultural landscapes. Department of Biology/Center for Environmental and Climate Research

Lina Nikoleris (2016) The estrogen receptor in fish and effects of estrogenic substances in the environment: ecological and evolutionary perspectives and societal awareness Department of Biology/Center for Environmental and Climate Research

Cecilia Hultin (2016) Estrogen receptor and multixenobiotic resistance genes in freshwater fish and snails: identification and expression analysis after pharmaceutical exposure. Center for Environmental and Climate Research

Annika M. E. Söderman (2016) Small biotopes: Landscape and management effects on pollinators. Department of Biology/Center for Environmental and Climate Research

Wenxin Ning (2016) Tracking environmental changes of the Baltic Sea coastal zone since the mid-Holocene. Department of Geology/Center for Environmental and Climate Research

Karin Mattsson (2016) Nanoparticles in the aquatic environment, Particle characterization and effects on organisms. Department of Chemistry/Center for Environmental and Climate Research

Ola Svahn (2016) Tillämpad miljöanalytisk kemi för monitorering och åtgärder av antibiotika- och läkemedelsrester I Vattenriket. School of Education and Environment, Kristianstad university

Pablo Urrutia Cordero (2016) Putting food web theory into action: Local adaptation of freshwaters to global environmental change. Department of Biology/Center for Environmental and Climate Research

Lin Yu (2016) Dynamic modelling of the forest ecosystem: Incorporation of the phosphorous cycle. Center for environmental and Climate Research

Behnaz Pirzamanbein (2016) Reconstruction of Past European Land Cover Based on Fossil Pollen Data: Gaussian Markov Random Field Models for Compositional Data. Center for Mathematical Sciences/Center for Environmental and Climate Research

Printed by Media-Tryck, Lund University 2016
Nordic Ecolabel 3041 0903

Behnaz Pirzamanbein works with Spatial statistics methods. Her PhD study is a multidisciplinary research between the Center for Mathematical Sciences and the Center for Environmental and climate research at Lund University.
This research aims at creating maps of land cover based on fossil pollen data for different time periods over the past 6000 years.
The model developed in this thesis was awarded for Outstanding Bayesian research applied to climate science in the Section on Bayesian Statistical Science (SBSS), and also recognized with an Honorary Mention in the Section on Statistics and the Environment (ENVR), in the student paper competition at the joint statistical meeting, ASA, 2016.

This thesis consists of the following papers:

Paper A Pirzamanbein B., Lindström J., Poska A., Sugita S., Trondman A., Fyfe R., Mazier F., Nielsen A., Kaplan J., Bjune A., Birks J., Giesecke T., Kangur M, Latałowa M., Marquer L., Smith B. and Gaillard M.: Creating spatially continuous maps of past land cover from point estimates: A new statistical approach applied to pollen data *Ecological Complexity*, 2014:12, 20, 127–141.

Paper B Pirzamanbein B., Lindström J., Poska A., and Marie-José Gaillard: Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties.
Published as preprint on Arxiv: arxiv.org/abs/1511.06417
To be submitted to *Journal of the American Statistical Association*.

Paper C Pirzamanbein B., Poska A., Lindström J.: Analysing the sensitivity of pollen based land cover maps to different auxiliary variables To be submitted to *Climate of Past*.

Paper D Pirzamanbein B., Lindström J.: Reconstruction of past Human Land Use from Pollen data and Anthropogenic Land Cover Changes scenarios To be submitted to *Environmetrics*.

LUND UNIVERSITY

Lund University
Faculty of Science
Centre for Matemathical Sciences
Centre for Environmental and Climate Research
ISBN 978-91-7753-076-3
ISSN 1404-0034