Bayesian reconstruction of past land-cover from pollen data: model robustness and sensitivity to auxiliary variables

Pirzamanbein, Behnaz; Poska, Anneli; Lindström, Johan

Published in:
Earth and Space Science

Link to article, DOI:
10.1029/2018ea000547

Publication date:
2019

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

Bayesian reconstruction of past land-cover from pollen data: model robustness and sensitivity to auxiliary variables

Behnaz Pirzamanbein[1,2,3], Anneli Poska[4,5], Johan Lindström[2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark

[2]Centre for Mathematical Sciences, Lund University, Sweden

[3]Centre for Environmental and Climate Research, Lund University, Sweden

[4]Department of Physical Geography and Ecosystems Analysis, Lund University, Sweden

[5]Institute of Geology, Tallinn University of Technology, Estonia

Key Points:

- Introduces a new set of North European pollen-proxy based land-cover reconstructions.
- Presents a spatial statistical interpolation model to create pollen-proxy based reconstructions.
- The method is stable even when using (very) different auxiliary datasets.

Corresponding author: Behnaz Pirzamanbein, bepi@dtu.dk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2018EA000547

Abstract

Realistic depictions of past land cover are needed to investigate prehistoric environmental changes, effects of anthropogenic deforestation, and long term land cover-climate feedbacks. Observation based reconstructions of past land cover are rare and commonly used model based reconstructions exhibit considerable differences. Recently Pirzamanbein, Lindström, Poska, and Gaillard (Spatial Statistics, 24:14–31, 2018) developed a statistical interpolation method that produces spatially complete reconstructions of past land cover from pollen assemblage. These reconstructions incorporate a number of auxiliary datasets raising questions regarding the method's sensitivity to different auxiliary datasets.

Here the sensitivity of the method is examined by performing spatial reconstructions for northern Europe during three time periods (1900 CE, 1725 CE and 4000 BCE). The auxiliary datasets considered include the most commonly utilized sources of past land-cover data — e.g. estimates produced by a dynamic vegetation (DVM) and anthropogenic land-cover change (ALCC) models. Five different auxiliary datasets were considered, including different climate data driving the DVM and different ALCC models. The resulting reconstructions were evaluated using cross-validation for all the time periods. For the recent time period, 1900 CE, the different land-cover reconstructions were also compared against a present day forest map.

The validation confirms that the statistical model provides a robust spatial interpolation tool with low sensitivity to differences in auxiliary data and high capacity to capture information in the pollen based proxy data. Further auxiliary data with high spatial detail improves model performance for areas with complex topography or few observations.

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, Stott, Hegerl, & Betts, 2013; Claussen, Brovkin, & Ganopolski, 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, especially for Arctic areas (Chapman & Walsh, 2007; Koenigk et al., 2013; Miller & Smith, 2012; Richter-Menge, Jeffries, & Overland,

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, Brandefelt, Kjellström, & Smith, 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, Reick, Raddatz, & Claussen, 2008; Strandberg et al., 2014). Differences in input climate variables (such as temperature, precipitation and etc. which may affect DVM output, see Wu et al., 2017, for details), mechanistic and parametrisation differences of DVMs (Prentice et al., 2007; Scheiter, Langan, & Higgins, 2013), and significant variation between existing ALCC scenarios (e.g. Gaillard et al., 2010; Goldewijk, Beusen, Van Drecht, & De Vos, 2011; Kaplan, Krumhardt, & Zimmermann, 2009; 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 (Pitman et al., 2009; Strandberg et al., 2014). Thus, reliable land-cover representations are important when studying biogeophysical impacts of anthropogenic land-cover change on climate.

The palaeoecological proxy based land-cover reconstructions recently published by Pirzamanbein et al. (2018, 2014) were designed to overcome the problems described above. And to provide a proxy 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 each of the studied sites. Spatial interpolation is needed to fill the gaps between observations and to produce continuous land-cover reconstructions. Conventional interpolation meth-

ods might struggle when handling noisy, spatially heterogeneous data (de Knegt et al., 2010; Heuvelink, Burrough, & Stein, 1989), but statistical methods for spatially structured data exist (Blangiardo & Cameletti, 2015; Gelfand, Diggle, Guttorp, & Fuentes, 2010).

In Pirzamanbein et al. (2018) a statistical model based on Gaussian Markov Random Fields (Lindgren, Rue, & Lindström, 2011; Rue & Held, 2005) was developed to provide a reliable, computationally effective and freeware based spatial interpolation technique. The resulting statistical model combines PbLCC data with auxiliary datasets; e.g. DVM output, ALCC scenarios, and elevation; to produce reconstructions of past land cover. The auxiliary data is subject to the differences and uncertainties outlined above and the choice of auxiliary data could influence the accuracy of the statistical model. The major objectives of this paper are: 1) To draw attention of climate modelling community to a novel set of spatially explicit pollen-proxy based land-cover reconstructions suitable for climate modelling; 2) to present and test the robustness of the spatial interpolation model developed by Pirzamanbein et al. (2018); and 3) to evaluate the models capacity to recover information provided by PbLCC proxy data and to analyse its sensitivity to different 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 PbLCC data published in Trondman et al. (2015) consists of proportions of coniferous forest, broadleaved forest and unforested land presented as gridded (1°×1°) data points placed irregularly across northern-central Europe. Altogether 175 grid cells containing proxy 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, along with elevation were considered as potential auxiliary datasets. In each case potential natural vegetation composition estimated by the DVM LPJ-GUESS (Lund-Potsdam-Jena General Ecosystem Simulator; Sitch et al., 2003; Smith, Prentice, & Sykes, 2001) were combined with an ALCC scenario to adjust for human land use (see Pirzamanbein et al., 2014, for details). The model derived estimates of the past land cover were obtained using DVM LPJ-GUESS simulated percentage cover of the plant functional types (PFTs) defined

for Europe by Hickler et al. (2012). The PFTs were averaged over the specific modelled period and aggregated to three land-cover types (LCTs), i.e. Coniferous forest, Broadleaved forest and Unforested land. The climate forcings used as an environmental driver in DVM were derived from two climate models: Earth System Model (ESM Mikolajewicz et al., 2007) and Rossby Centre Regional Climate Model (RCA3 Samuelsson et al., 2011). Since anthropogenic deforestation and human land-use is not accounted for by LPJ-GUESS, ALCC data derived from the two most commonly used ALCC scenarios: the standard KK10 scenario by Kaplan et al. (2009) and the History Database of the Global Environment (HYDE) scenario by Goldewijk et al. (2011). The human land-use data was used to adjust the LCT estimates by decreasing the proportion of all three LCT fractions by the human land-use fraction, thereafter the human land-use fraction was added to the Unforested land fraction.

K-L$_{\mathbf{RCA3}}$: Combines the ALCC scenario KK10 (Kaplan et al., 2009) and the potential natural vegetation from LPJ-GUESS. Climate forcing for the DVM was derived from Rossby Centre Regional Climate Model (RCA3, Samuelsson et al., 2011) at annual time and $0.44° \times 0.44°$ spatial resolution (Figure 1, column 3),

K-L$_{\mathbf{ESM}}$: Combines the ALCC scenario KK10 and the potential natural vegetation from LPJ-GUESS. 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 was used (Figure 1, column 4). This additional data provides information of the observed climate variability at the temporal and spatial scales during the interpolation,

H-L$_{\mathbf{RCA3}}$: Combines the ALCC scenario from the History Database of the Global Environment (HYDE; Goldewijk et al., 2011) and vegetation from LPJ-GUESS with RCA3 climate forcing (Figure 1, column 5),

H-L$_{\mathbf{ESM}}$: Combines the ALCC scenario from HYDE and vegetation from LPJ-GUESS with ESM climate forcing (Figure 1, column 6).

The elevation data (denoted SRTM$_{\mathrm{elev}}$) was obtained from the Shuttle Radar Topography Mission (Becker et al., 2009) (Figure 1, column 1 row 2).

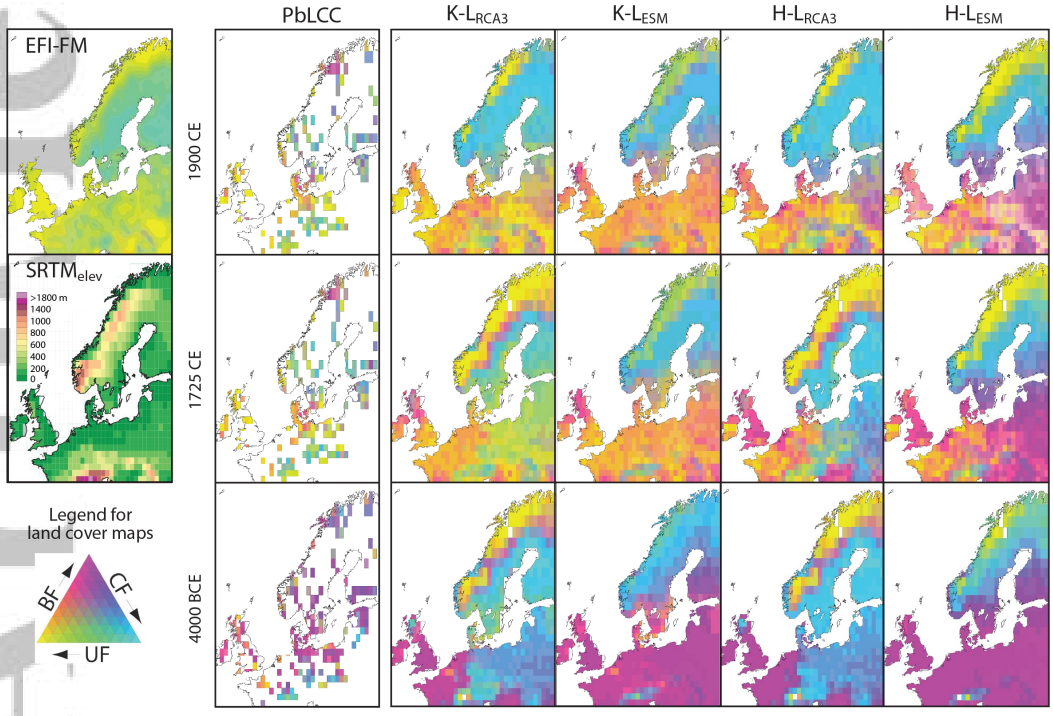

Figure 1. Data used in the modelling. The first column shows (from top to bottom) the EFI-FM, $SRTM_{elev}$, and the colorkey for the land-cover compositions, coniferous forest (CF), broadleaved forest (BF) and unforested land (UF). The remaining columns give (from left to right) the PbLCC (Trondman et al., 2015) and the four model based compositions considered as potential covariates: K-L_{RCA3}, K-L_{ESM}, H-L_{RCA3}, and H-L_{ESM}. Here K/H indicates KK10 (Kaplan et al., 2009) or HYDE (Goldewijk et al., 2011) land use scenarios and L_{RCA3}/L_{ESM} indicates the climate — Rossby Centre Regional Climate Model (Samuelsson et al., 2011) or Earth System Model (Mikolajewicz et al., 2007) — used to drive the vegetation model. The three rows represent (from top to bottom) the time periods 1900 CE, 1725 CE, and 4000 BCE.

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 and national forest-inventory statistics from 1990–2005 (Pivinen et al., 2001; Schuck et al., 2002) (Figure 1, column 1 row 1). All auxiliary data were up-scaled to $1° \times 1°$ spatial resolution, matching the pollen based reconstructions, before usage as model input.

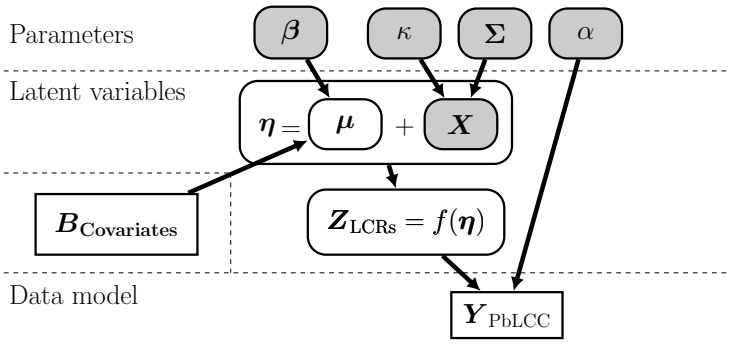

Figure 2. Hierarchical graph describing the conditional dependencies between observations (white rectangle) and parameters (grey rounded rectangles) to be estimated. The white rounded rectangles are computed based on the estimations. In a generalized linear mixed model framework, $\boldsymbol{\eta}$ is the linear predictor — consisting of a regression term, $\boldsymbol{\mu}$, and a spatial random effect, $\boldsymbol{X}$. The link function, $f(\boldsymbol{\eta})$, transforms between linear predictor and proportions, which are matched to the observed land cover proportions, $\boldsymbol{Y}_{\text{PbLCC}}$, using a Dirichlet distribution.

2.1 Statistical Model for Land-cover Compositions

A Bayesian hierarchical model is used to interpolate the PbLCC data; here we only provide a brief overview of the model, mathematical and technical details can be found in Pirzamanbein et al. (2018). The model can be seen as a special case of a generalized linear mixed model with a spatially correlated random effect. An alternative interpretation of the model is as an empirical forward model (direction of arrows in Figure 2) where parameters affect the latent variables which in turn affect the data. Reconstructions are obtained by inverting the model (i.e. computing the posterior) to obtain the latent variables given the data.

The PbLLC derived proportions of land cover (coniferous forest, broadleaved forest and unforested land), denoted $\boldsymbol{Y}_{\text{PbLCC}}$, are seen as draws from a Dirichlet distribution (Kotz, Balakrishnan, & Johnson, 2000, Ch. 49) given a vector of proportions, $\boldsymbol{Z}$, and a concentration parameter, α (controlling the uncertainty: $\mathsf{V}(\boldsymbol{Y}_{\text{PbLCC}}) \propto 1/\alpha$). Since the proportions have to obey certain restrictions ($0 \leq Z_k \leq 1$ and $\sum_{k=1}^{3} Z_k = 1$, were k indexes the land-cover types), a link function is used to transform between the pro-

portions and the linear predictor, $\boldsymbol{\eta}$:

$$Z_k = f(\boldsymbol{\eta}) = \begin{cases} \frac{e^{\eta_k}}{1+\sum_{i=1}^{2} e^{\eta_i}} & \text{for } k = 1, 2 \\ \frac{1}{1+\sum_{i=1}^{2} e^{\eta_i}} & \text{for } k = 3 \end{cases}$$

$$\eta_k = f^{-1}(\boldsymbol{Z}) = \log\left(\frac{Z_k}{Z_3}\right) \qquad \text{for } k = 1, 2$$

Here $f^{-1}(\boldsymbol{Z})$ is the additive log-ratio transformation (Aitchison, 1986), a multivariate extension of the logit transformation.

The linear predictor consists of a mean structure and a spatially dependent random effect, $\boldsymbol{\eta} = \boldsymbol{\mu} + \boldsymbol{X}$. The mean structure is modelled as a linear regression, $\boldsymbol{\mu} = \boldsymbol{B\beta}$; i.e. a combination of covariates, $\boldsymbol{B}$, and regression coefficients, $\boldsymbol{\beta}$. To aid in variable selection and suppress uninformative covariates a horseshoe prior (Makalic & Schmidt, 2016; Park & Casella, 2008) is used for $\boldsymbol{\beta}$. The main focus of this paper is to evaluate the model sensitivity to the choice of covariates (i.e. the auxiliary datasets). The PbLCC is modelled based on six different sets of covariates: 1) Intercept, 2) $\text{SRTM}_{\text{elev}}$, 3) K-L_{ESM}, 4) K-L_{RCA3}, 5) H-L_{ESM}, and 6) H-L_{RCA3}; illustrated in Figure 1. A summary of the different models is given in Table 1. Finally, the spatially dependent random effect,$\boldsymbol{X}$, is modelled using a Gaussian Markov Random Field (Lindgren et al., 2011) with two parameters: κ, controlling the strength of the spatial dependence and $\boldsymbol{\Sigma}$, controlling the variation within and between the fields (i.e. the correlation among different land-cover types).

Model estimation and reconstructions are performed using Markov Chain Monte Carlo (Brooks, Gelman, Jones, & Meng, 2011) with 100 000 samples and a burn-in of 10 000 (See Pirzamanbein et al., 2018, for details.). Output from the Markov Chain Monte Carlo are then used to compute land-cover reconstructions (as posterior expectations, $\mathsf{E}(\boldsymbol{Z}|\boldsymbol{Y}_{\text{PbLCC}})$) and uncertainties in the form of predictive regions. The predictive regions describe the uncertainties associated with the reconstructions; including uncertainties in model parameters and linear predictor.

2.2 Testing the Model Performance

To evaluate model performance, we compared the land-cover reconstructions from different models for the 1900 CE time period with the EFI-FM by computing the average compositional distances (ACD; Aitchison, Barceló-Vidal, Martín-Fernández, & Pawlowsky-

Table 1. Six different models and corresponding covariates. $SRTM_{elev}$ is elevation (Becker et al., 2009), K/H indicates KK10 (Kaplan et al., 2009) or HYDE (Goldewijk et al., 2011) land use scenarios and L_{RCA3}/L_{ESM} indicates vegetation model driven by climate from the Rossby Centre Regional Climate Model (Samuelsson et al., 2011) or from an Earth System Model (Mikolajewicz et al., 2007).

Model \ Covariates	Intercept	$SRTM_{elev}$	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

Glahn, 2000; Pirzamanbein, 2016; Pirzamanbein et al., 2018). This measure is similar to root mean square error in $\mathbb{R}^2$ but it accounts for compositional properties (i.e. $0 \leq Z_k \leq 1$ and $\sum_{k=1}^{3} Z_k = 1$) and it is computed by

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

where u and v are additive log-ratio transforms of the compositions to be compared and $J_{k-1\times k-1}$ is a matrix with elements $J_{l,l} = 2$ and $J_{l,p} = 1$ which neutralizes the choice of denominator in alr transformation.

Since no independent observational data exists for the 1725 CE and 4000 BCE time periods, we applied a 6-fold cross-validation scheme (Hastie, Tibshirani, & Friedman, 2001, Ch. 7.10) to all models and time periods. The cross-validation divides the observations into 6 random groups and the reconstruction errors for each group when using only observations from the other 5 groups are computed. To further compare predictive performance of the models Deviance Information Criteria (DIC; see Ch. 7.2 in Gelman et al., 2014) were computed for all models and time periods. The DIC is a hierarchical modelling generalization of the Akaike and Bayesian information criteria (Hastie et al., 2001, Ch. 7).

3 Results and Discussion

Fossil pollen is a well-recognized information source of vegetation dynamics and generally accepted as the best observational data on past land-cover composition and environmental conditions (Trondman et al., 2015).

Today, central and northern Europe have, at the subcontinental spatial scale, the highest density of palynologically investigated sites on Earth. However, even there the existing pollen records are irregularly placed, leaving some areas with scarce data coverage (Fyfe, Woodbridge, & Roberts, 2015). The collection of new pollen data to fill these gaps is very time consuming and cannot be performed everywhere. All this makes pollen data, in their original format, heavily underused, since the data is unsuitable for models requiring continuous land-cover representations as input. The lack of spatially explicit proxy based land cover data directly usable in climate models has been hampering the correct representation of past climate-land cover relationship.

Regrettably, the commonly used DVM derived representations of past land cover exhibit large variation in vegetation composition estimates. The model derived land-cover datasets used as auxiliary data (Table 1) show large variation in estimated extents of coniferous and broadleaved forests, and unforested areas for all of the studied time periods (Figure 1). These substantial differences illustrate large deviances between model based estimates of the past land-cover composition due to differences in applied climate forcing and/or ALCC scenarios. Differences in climate model outputs (Gladstone et al., 2005; Harrison et al., 2014) and ALCC model estimates (Gaillard et al., 2010) have been recognized in earlier comparison studies and syntheses. The effect of the differences in input climate forcing and ALCC scenario on DVM estimated land-cover composition presented here are especially pronounced for central and western Europe, and for elevated areas in northern Scandinavia and the Alps (Figure 1). In general the KK10 ALCC scenario produces larger unforested areas, notably in western Europe, compared to the HYDE scenario. Compared to the ESM climate forcing; the RCA3 forcing results in higher proportions of coniferous forest, especially for central, northern and eastern Europe. The described differences are clearly recognizable for all the considered time periods and are generally larger between time periods than within each time period. The purpose of the statistical model presented in Section 2.1 is to combine the observed PbLCC with the

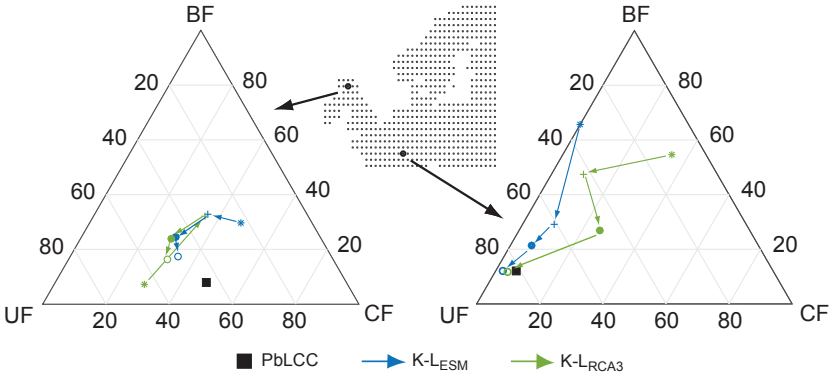

Figure 3. Advancement of the model for two locations at 1725 CE. Starting from the value of the K-L_{RCA3} and K-L_{ESM} covariates ($*$), the cumulative effects of regression coefficients, $\boldsymbol{\beta}$, ($+$); the intercept and $SRTM_{elev}$ covariates ($\bullet$); and, finally, the spatial dependency structures ($\circ$), are illustrated. With the final points ($\circ$) corresponding to the land-cover reconstructions and ■ marking the observed pollen based land-cover composition.

spatial structure in the auxiliary data to produce data driven spatially complete maps of past land-cover that can be used directly (as input) in others models.

To illustrate the structure of the statistical model, step by step advancement from auxiliary data (model derived land cover) to final statistical estimates of land-cover compositions for 1725 CE are given in Figures 3 and 4. The large differences in K-L_{RCA3} and K-L_{ESM} are reduced by scaling with the regression coefficients, $\boldsymbol{\beta}$, capturing the empirical relationship between covariates and PbLCC data. Thereafter, the land-cover estimates are subjected to similar adjustments due to intercept and $SRTM_{elev}$, and finally similar spatial dependent effects.

The impact of different auxiliary datasets was assessed by using the statistical model to create a set of proxy based reconstructions of past land cover for central and northern Europe during three time periods (1900 CE, 1725 CE and 4000 BCE; see Figures 5 and 6). Each of these reconstructions were based on the irregularly distributed observed pollen data (PbLCC), available for ca 25% of the area, together with one of the six models (Table 1) using different combinations of the auxiliary data (Figure 1).

The resulting land-cover reconstructions exhibit considerably higher similarity with the PbLCC data than any of the auxiliary land-cover datasets for all tested models and time periods (Figures 5 and 6). At first the similarity among the reconstructions might

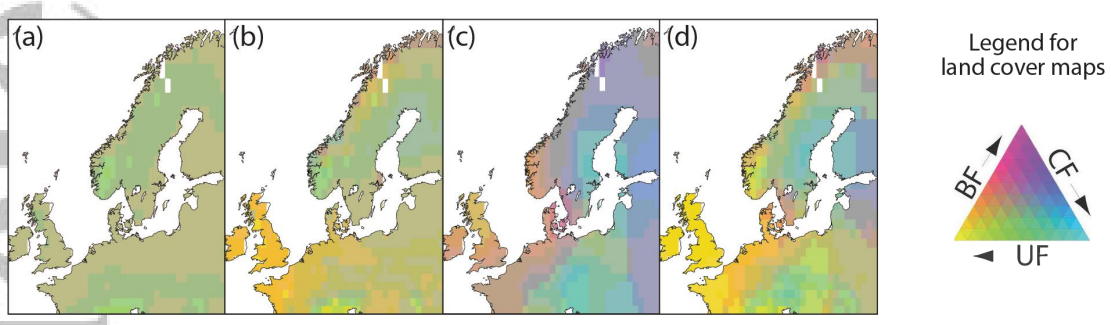

Figure 4. Advancement of K-L$_{\text{ESM}}$ models for the 1725 CE time period: (a) shows the effect of intercept and SRTM$_{\text{elev}}$, (b) shows the mean structure, $\boldsymbol{\mu}$, including all the covariates, (c) shows the spatial dependency structure and finally (d) shows the resulting land-cover reconstructions obtained by adding (b) and (c).

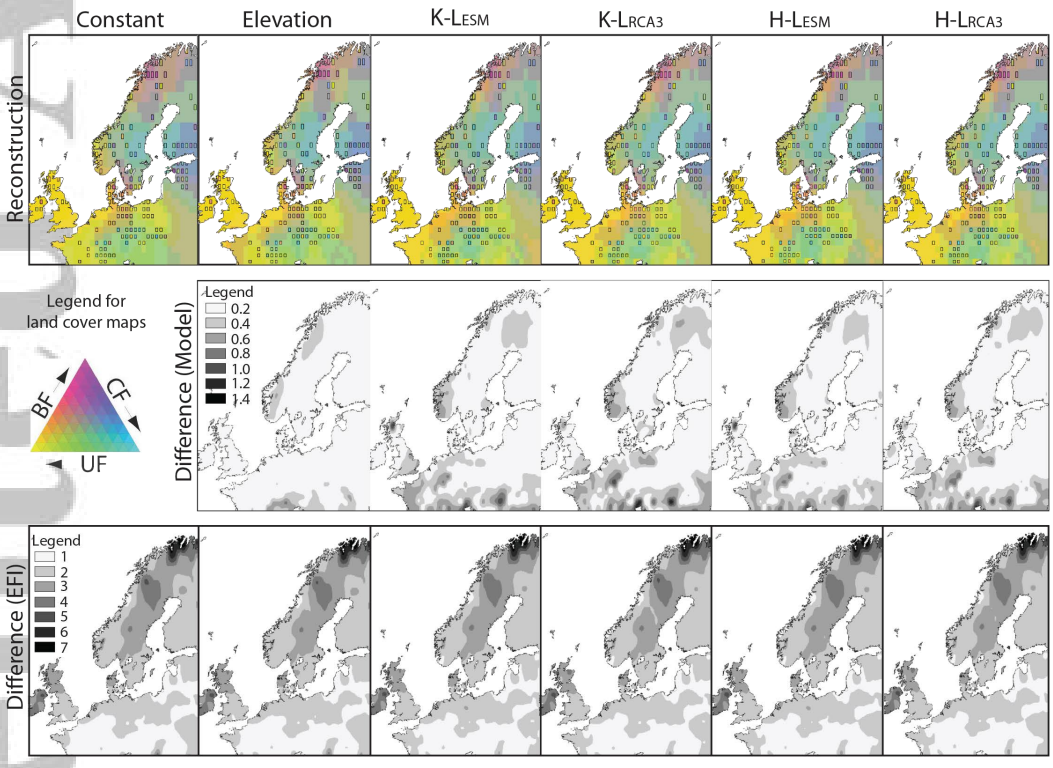

Figure 5. Land-cover reconstructions using PbLCC for the 1900 CE time periods (top row). The reconstructions are based on six different models (see Table 1) with different auxiliary datasets. Locations and compositional values of the available PbLCC data are given by the black rectangles; these rectangles match the locations of available data as illustrated in column 2 of Figure 1. Middle row shows the compositional distances between each model and the Constant model. Bottom row shows the compositional distances between each model and the EFI-FM.

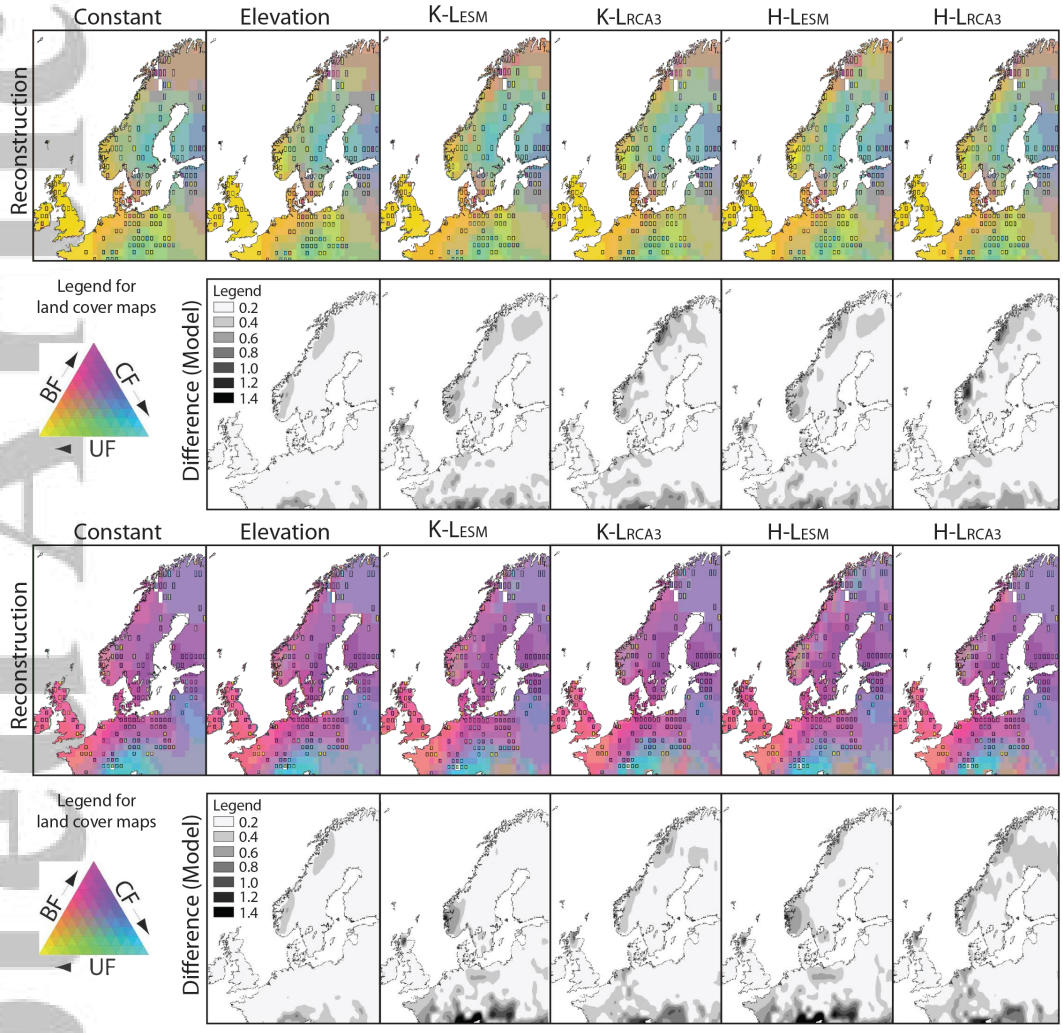

Figure 6. Land-cover reconstructions using local estimates of PbLCC for the 1725 CE (top) and 4000 BCE (bottom) time periods. The reconstructions are based on six different models (see Table 1) with different auxiliary datasets. Locations and compositional values of the available PbLCC data are given by the black rectangles; these rectangles match the locations of available data as illustrated in column 2 of Figure 1. Second and fourth row show the compositional distances between each model and the Constant model.

Table 2. Deviance information criteria (DIC) and Average compositional distances (ACD) from 6-fold cross-validations for each of the six models and three time periods. Best value for each time period marked in **bold-font**.

	DIC			ACD		
	1900CE	1725CE	4000BCE	1900CE	1725CE	4000BCE
Constant	-559	-655	-593	1.00	1.12	1.20
Elevation	**-568**	-664	-589	0.99	**1.11**	1.21
K-L_{ESM}	-547	-649	**-609**	1.00	1.12	1.18
K-L_{RCA3}	-549	-661	-589	0.99	1.13	1.19
H-L_{ESM}	-549	-655	-608	0.99	1.11	**1.17**
H-L_{RCA3}	-557	**-669**	-595	**0.99**	1.12	1.18

seem contradictory, but recall that the model allows for, and estimates, different weighting (the regression coefficients, $\boldsymbol{\beta}$:s) for each of the auxiliary datasets. Thus, the resulting reconstructions do not rely on the absolute values in the auxiliary datasets, only their spatial patterns. As a result, model performance for elevated areas and for the areas with low observational data coverage (e.g. eastern and south-eastern Europe) is improved by including covariates that exhibit distinct spatial structures for the given areas (Figures 5 and 6). Neither the DIC results nor the 6-fold cross-validation results show any advantage among the six tested models for the different time periods (Table 2). The DIC values share many properties of AIC values, and as pointed out in Burnham and Anderson (1998) models within 2 units of the *best* are equivalent, within 3-7 have less support and models differing more than 10 units are essentially unsupported. For the CV values a simulation study in Pirzamanbein et al. (2018) indicates that the standard deviation due to the random ordering of validation points is around 0.01 units, and thus models differing by less than 0.02 (2σ) could be considered equivalent. Analogous to the reconstructions, the predictive regions are very similar in both size and shape irrespective of the auxiliary dataset used, indicating similar reconstruction uncertainties across all models (Figure 7). Implying there is no clear preference among the models, i.e. that the results are robust to the choice of auxiliary dataset.

Although a temporal misalignment exists between the PbLCC data for the 1900 CE time period (based on pollen data from 1850 to the present) and the EFI-FM (in-

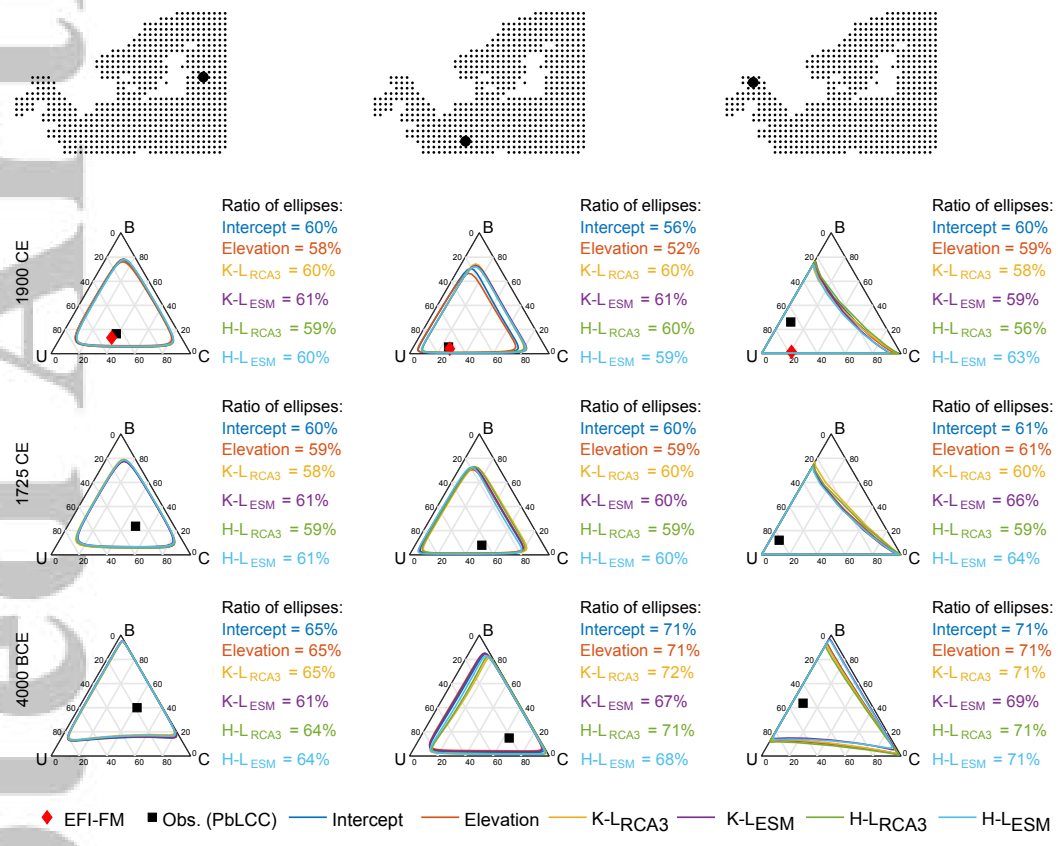

Figure 7. The prediction regions and fraction of the ternary triangle covered by these regions are presented for three locations, the six models, and the 1900 CE, 1725 CE and 4000 BCE time periods.

ventory and satellite data from 1990-2005); EFI-FM provides the best complete and consistent land-cover map of Europe for present time, making it a reasonable choice for a comparison. The main differences between the EFI-FM and the PbLCC data for the 1900 CE time period are: 1) lower abundance of broadleaved forests for most of Europe, 2) higher abundance of coniferous forest in Sweden and Finland, and 3) higher abundance of unforested land in North Norway in the EFI-FM data than in the PbLCC data (Pirzamanbein et al., 2018). The average compositional distances computed between the land-cover reconstructions and the EFI-FM for 1900 CE show practically identical (1.47 to 1.48) distances between all six reconstructions and the EFI-FM, and small differences among the six presented models (Table 3).

These results clearly show that the developed statistical interpolation model is robust to the choice of covariates. The model is suitable for reconstructing spatially continuous maps of past land cover from scattered and irregularly spaced pollen based proxy data.

4 Conclusions

The statistical model and Bayesian interpolation method presented here has been specially designed for handling irregularly spaced palaeo-proxy records like pollen data and, dependent on proxy data availability, is globally applicable. The model produces land-cover maps by combining irregularly distributed pollen based estimates of land cover with auxiliary data and a statistical model for spatial structure. The resulting maps capture important features in the pollen proxy data and are reasonably insensitive to the use of different auxiliary datasets.

Auxiliary datasets considered were compiled from commonly utilized sources of past land-cover data (outputs from a dynamic vegetation model using different climatic drivers and anthropogenic land-cover changes scenarios). These datasets exhibit considerable differences in their recreation of the past land cover. Emphasizing the need for the independent, proxy based past land-cover maps created in this paper.

Evaluation of the model's sensitivity indicates that the proposed statistical model is robust to the choice of auxiliary data and only considers features in the auxiliary data that are consistent with the proxy data. However, auxiliary data with detailed spatial

Table 3. The average compositional distances among the six models fitted to the data for each of the three time periods.

	EFI-FM	Elevation	K-L_{ESM}	K-L_{RCA3}	H-L_{ESM}	H-L_{RCA3}
	1900 CE					
Constant	1.48	0.08	0.18	0.20	0.17	0.19
Elevation	1.49		0.19	0.21	0.18	0.20
K-L_{ESM}	1.48			0.09	0.07	0.09
K-L_{RCA3}	1.48				0.11	0.06
H-L_{ESM}	1.48					0.08
H-L_{RCA3}	1.48					
	1725 CE					
Constant		0.10	0.16	0.16	0.17	0.17
Elevation			0.14	0.11	0.14	0.13
K-L_{ESM}				0.14	0.06	0.16
K-L_{RCA3}					0.15	0.07
H-L_{ESM}						0.15
	4000 BCE					
Constant		0.11	0.21	0.17	0.22	0.19
Elevation			0.19	0.12	0.20	0.15
K-L_{ESM}				0.19	0.07	0.21
K-L_{RCA3}					0.18	0.07
H-L_{ESM}						0.20

information considerably improves the interpolation results for areas with low proxy data coverage, with no reduction in overall performance.

This modelling approach has demonstrated a clear capacity to produce empirically based land-cover reconstructions for climate modelling purposes. Such reconstructions are necessary to evaluate anthropogenic land-cover change scenarios currently used in climate modelling and to study past interactions between land cover and climate with greater reliability. The model will also be very useful for producing reconstructions of past land cover from the global pollen proxy data currently being produced by the PAGES (Past Global changES) LandCover6k initiative.

5 Data availability

The database containing the reconstructions of coniferous forest, broadleaved forest and unforested land, three fractions of land cover, for the three time-periods presented in this paper, along with reconstructions for 1425 CE and 1000 BCE using only the K-L_{ESM} are available for download from `https://github.com/BehnazP/SpatioCompo`. The PbLCC data is available from `https://doi.pangaea.de/10.1594/PANGAEA.897303`.

Acronyms

DVM Dynamical vegetation model.

ALCC Anthropogenic land-cover change.

PbLCC Pollen based land-cover composition.

LPJ-GUESS The Lund-Potsdam-Jena General Ecosystem Simulator, a DVM.

EFI-FM European Forest Institute forest map.

Notation

$\boldsymbol{Y}_{\mathbf{PbLCC}}$ Observations, as proportions.

$\boldsymbol{f}$ Link function, transforming between proportions and linear predictor.

$\boldsymbol{\eta}$ Linear predictor, $\boldsymbol{\eta} = \boldsymbol{\mu} + \boldsymbol{X}$.

$\boldsymbol{\mu}$ Mean structure; modelled as $\boldsymbol{\mu} = \boldsymbol{B\beta}$ using covariates, $\boldsymbol{B}$, and regression coefficients, $\boldsymbol{\beta}$.

$\boldsymbol{X}$ Spatially dependent random effect.

$\boldsymbol{\alpha}$ Concentrated parameter of the Dirichlet distribution (i.e. observational uncertainty)

320 $\mathbf{\Sigma}$ Covariance matrix that determines the variation between and within fields

321 $\boldsymbol{\kappa}$ Scale parameter controlling the range of spatial dependency

322 **Acknowledgments**

323 The research presented in this paper is a contribution to the two Swedish strategic re-
324 search areas Biodiversity and Ecosystems in a Changing Climate (BECC), and ModElling
325 the Regional and Global Earth system (MERGE). The paper is also a contribution to
326 PAGES LandCover6k. Lindström has been funded by Swedish Research Council (SRC,
327 Vetenskapsrådet) grant no 2012-5983. Poska has been funded by SRC grant no 2016-03617
328 and the Estonian Ministry of Education grant IUT1-8. The authors would like to acknowl-
329 edge Marie-José Gaillard for her efforts in providing the pollen based land-cover proxy
330 data and thank her for valuable comments on this manuscript.

331 **References**

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

334 Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J., & Pawlowsky-Glahn, V.
335 (2000). Logratio analysis and compositional distance. *Math. Geol.*, *32*(3),
336 271–275.

337 Arneth, A., Harrison, S. P., Zaehle, S., Tsigaridis, K., Menon, S., Bartlein, P. J., ...
338 others (2010). Terrestrial biogeochemical feedbacks in the climate system.
339 *Nature Geosci.*, *3*(8), 525–532. doi: 10.1038/ngeo905

340 Becker, J. J., Sandwell, D. T., Smith, W. H. F., Braud, J., Binder, B., Depner, J.,
341 ... Weatherall, P. (2009). Global bathymetry and elevation data at 30 arc
342 seconds resolution: SRTM30_PLUS. *Marine Geol.*, *32*(4), 355–371.

343 Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models*
344 *with r-inla.* Wiley.

345 Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). *Handbook of Markov*
346 *Chain Monte Carlo.* CRC Press.

347 Brovkin, V., Bendtsen, J., Claussen, M., Ganopolski, A., Kubatzki, C., Petoukhov,
348 V., & Andreev, A. (2002). Carbon cycle, vegetation, and climate dynamics in
349 the holocene: Experiments with the CLIMBER-2 model. *Glob. Biogeochem.*

Cycles, *16*(4), 1139.

Brovkin, V., Claussen, M., Driesschaert, E., Fichefet, T., Kicklighter, D., Loutre, M., ... Sokolov, A. (2006). Biogeophysical effects of historical land cover changes simulated by six Earth system models of intermediate complexity. *Clim. Dyn.*, *26*(6), 587–600. doi: 10.1007/s00382-005-0092-6

Burnham, K. P., & Anderson, D. R. (1998). Practical use of the information-theoretic approach. In *Model selection and inference* (pp. 75–117). Springer.

Chapman, W. L., & Walsh, J. E. (2007). Simulations of Arctic temperature and pressure by global coupled models. *J. Clim.*, *20*(4), 609–632. doi: 10.1175/JCLI4026.1

Christidis, N., Stott, P. A., Hegerl, G. C., & Betts, R. A. (2013). The role of land use change in the recent warming of daily extreme temperatures. *Geophys. Res. Lett.*, *40*(3), 589–594. doi: 10.1002/grl.50159

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

de Knegt, H. J., van Langevelde, F., Coughenour, M. B., Skidmore, A. K., de Boer, W. F., Heitkönig, I. M. A., ... Prins, H. H. T. (2010). Spatial autocorrelation and the scaling of species–environment relationships. *Ecology*, *91*(8), 2455–2465. doi: 10.1890/09-1359.1

de Noblet-Ducoudré, N., Boisier, J.-P., Pitman, A., Bonan, G., Brovkin, V., Cruz, F., ... Voldoire, A. (2012). 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. Clim.*, *25*(9), 3261–3281. doi: 10.1175/JCLI-D-11-00338.1

Fyfe, R. M., Woodbridge, J., & Roberts, N. (2015). From forest to farmland: pollen-inferred land cover change across Europe using the pseudobiomization approach. *Glob. Change Biol.*, *21*(3), 1197–1212. doi: 10.1111/gcb.12776

Gaillard, M.-J., Sugita, S., Mazier, F., Trondman, A.-K., Brostrom, A., Hickler, T., ... Seppä, H. (2010). Holocene land-cover reconstructions for studies on land cover-climate feedbacks. *Clim. Past*, *6*, 483–499.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third ed.). Chapman & Hall/CRC.

Gladstone, R. M., Ross, I., Valdes, P. J., Abe-Ouchi, A., Braconnot, P., Brewer, S., ... G., V. (2005). Mid-Holocene NAO: A PMIP2 model intercomparison. *Geophys. Res. Lett.*, *32*(16), L16707. doi: 10.1029/2005GL023596

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

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., ... Willis, K. (2014). Climate model benchmarking with glacial and mid-Holocene climates. *Clim. Dyn.*, *43*(3–4), 671–688. doi: 10.1007/s00382-013-1922-6

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York, NY, USA: Springer New York Inc.

Heuvelink, G. B. M., Burrough, P. A., & Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *Int. J. Geogr. Inf. Syst.*, *3*(4), 303–322. doi: 10 .1080/02693798908941518

Hickler, T., Vohland, K., Feehan, J., Miller, P. A., Smith, B., Costa, L., ... Sykes, M. T. (2012). Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Glob. Ecol. Biogeogr.*, *21*(1), 50–63. doi: 10.1111/j.1466-8238.2010.00613.x

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

Koenigk, T., Brodeau, L., Graversen, R. G., Karlsson, J., Svensson, G., Tjernström, M., ... Wyser, K. (2013). Arctic climate change in 21st century CMIP5 simulations with EC-Earth. *Clim. Dyn.*, *40*(11-12), 2719–2743. doi: 10.1007/s00382-012-1505-y

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions. volume 1: Models and applications.* Wiley.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. B*, *73*(4), 423–498. doi: 10.1111/j.1467-9868.2011.00777.x

Makalic, E., & Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Lett.*, *23*(1), 179-182. doi: 10.1109/LSP.2015.2503725

Mikolajewicz, U., Gröger, M., Maier-Reimer, E., Schurgers, G., Vizcaíno, M., & Winguth, A. M. (2007). Long-term effects of anthropogenic CO2 emissions simulated with a complex earth system model. *Clim. Dyn.*, *28*(6), 599–633. doi: 10.1007/s00382-006-0204-y

Miller, P. A., & Smith, B. (2012). Modelling tundra vegetation response to recent arctic warming. *Ambio*, *41*(3), 281–291. doi: 10.1007/s13280-012-0306-1

Olofsson, J. (2013). *The Earth: climate and anthropogenic interactions in a long time perspective* (Doctoral dissertation, Lund University). Retrieved from `http://lup.lub.lu.se/record/3732052`

Park, T., & Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.*, *103*(482), 681–686. doi: 10.1198/016214508000000337

Pirzamanbein, B. (2016). *Reconstruction of past european land cover based on fossil pollen data Gaussian Markov random field models for compositional data* (Doctoral dissertation, Lund University). Retrieved from `http://lup.lub.lu.se/record/c2980af3-a480-45be-a346-80a33a8dd315` (ISBN 978–91–7753–076–3)

Pirzamanbein, B., Lindström, J., Poska, A., & Gaillard, M.-J. (2018). Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *Spatial Stat.*, *24*, 14–31. doi: 10.1016/j.spasta.2018.03.005

Pirzamanbein, B., Lindström, J., Poska, A., Sugita, S., Trondman, A.-K., Fyfe, R., ... Gaillard, M.-J. (2014). Creating spatially continuous maps of past land cover from point estimates: A new statistical approach applied to pollen data. *Ecol. Complex.*, *20*, 127–141. doi: 10.1016/j.ecocom.2014.09.005

Pitman, A., de Noblet-Ducoudré, N., Cruz, F., Davin, E., Bonan, G., Brovkin, V., ... Voldoire, A. (2009). Uncertainties in climate responses to past land cover change: First results from the LUCID intercomparison study. *Geophys. Res. Lett.*, *36*(14), n/a–n/a. doi: 10.1029/2009GL039076

Pivinen, R., Lehikoinen, M., Schuck, A., Hme, T., Vtinen, S., Kennedy, P., & Folving, S. (2001). *Combining Earth observation data and forest statistics* (Tech. Rep. No. 14). Joint Research Centre-European Commission.: European Forest Institute. Retrieved from `https://www.efi.int/publications-bank/`

combining-earth-observation-data-and-forest-statistics (ISBN: 952-9844-84-0 ISSN: 1238-8785)

Pongratz, J., Reick, C., Raddatz, T., & Claussen, M. (2008). A reconstruction of global agricultural areas and land cover for the last millennium. *Glob. Biogeochem. Cycles*, *22*(3), GB3018. doi: 10.1029/2007GB003153

Prentice, I. C., Bondeau, A., Cramer, W., Harrison, S. P., Hickler, T., Lucht, W., ... Sykes, M. T. (2007). Dynamic global vegetation modeling: quantifying terrestrial ecosystem responses to large-scale environmental change. In J. G. Canadell, D. E. Pataki, & L. F. Pitelka (Eds.), *Terrestrial ecosystems in a changing world. global change — the igbp series* (pp. 175–192). Springer. doi: 10.1007/978-3-540-32730-1_15

Richter-Menge, J. A., Jeffries, M. O., & Overland, J. E. (Eds.). (2011). *Arctic report card 2011.* National Oceanic and Atmospheric Administration. Retrieved from www.arctic.noaa.gov/reportcard

Rue, H., & Held, L. (2005). *Gaussian Markov random fields; theory and applications* (Vol. 104). Chapman & Hall/CRC.

Samuelsson, P., Jones, C. G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U., ... Wyser, K. (2011). The Rossby Centre regional climate model RCA3: model description and performance. *Tellus A*, *63*(1), 4–23.

Scheiter, S., Langan, L., & Higgins, S. I. (2013). Next-generation dynamic global vegetation models: learning from community ecology. *New Phytologist*, *198*(3), 957–969. doi: 10.1111/nph.12210

Schuck, A., van Brusselen, J., Päivinen, R., Häme, T., Kennedy, P., & Folving, S. (2002). *Compilation of a calibrated European forest map derived from NOAA-AVHRR data* (EFI Internal Report No. 13). EuroForIns.

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

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

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

Strandberg, G., Kjellström, E., Poska, A., Wagner, S., Gaillard, M.-J., Trondman, A.-K., ... Sugita, S. (2014). 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. Retrieved from `http://www.clim-past.net/10/661/2014/` doi: 10.5194/cp-10-661-2014

Trondman, A.-K., Gaillard, M.-J., Sugita, S., Mazier, F., Fyfe, R., Lechterbeck, J., ... Wick, L. (2015). Pollen-based quantitative reconstructions of past land-cover in NW Europe between 6k years BP and present for climate modelling. *Glob. Change Biol.*, *21*(2), 676–697. doi: 10.1111/gcb.12737

Wu, Z., Ahlström, A., Smith, B., Ardö, J., Eklundh, L., Fensholt, R., & Lehsten, V. (2017). Climate data induced uncertainty in model-based estimations of terrestrial primary productivity. *Environ. Res. Lett.*, *12*(6), 064013.

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