

**Comparing different calibration methods (WA/WA-PLS regression and Bayesian modelling) and different-sized calibration sets in pollen-based quantitative climate reconstruction**

J Sakari Salonen, Liisa Ilvonen, Heikki Seppä, Lasse Holmström, Richard J Telford, Andrejus Gaidamavicius, Migle Stancikaite and Dmitry Subetto

*The Holocene* 2012 22: 413 originally published online 30 November 2011

DOI: 10.1177/0959683611425548

The online version of this article can be found at:

<http://hol.sagepub.com/content/22/4/413>

---

Published by:



<http://www.sagepublications.com>

**Additional services and information for *The Holocene* can be found at:**

**Email Alerts:** <http://hol.sagepub.com/cgi/alerts>

**Subscriptions:** <http://hol.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

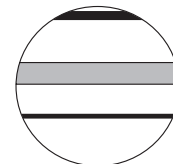
**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

**Citations:** <http://hol.sagepub.com/content/22/4/413.refs.html>

>> [Version of Record](#) - Mar 14, 2012

[OnlineFirst Version of Record](#) - Nov 30, 2011

[What is This?](#)



# Comparing different calibration methods (WA/WA-PLS regression and Bayesian modelling) and different-sized calibration sets in pollen-based quantitative climate reconstruction

J Sakari Salonen,<sup>1</sup> Liisa Ilvonen,<sup>2</sup> Heikki Seppä,<sup>1</sup>  
Lasse Holmström,<sup>2</sup> Richard J Telford,<sup>3</sup>  
Andrejus Gaidamavičius,<sup>4</sup> Miglė Stančikaitė<sup>4</sup> and Dmitry Subetto<sup>5</sup>

## Abstract

We compare a Bayesian modelling-based technique with weighted averaging (WA) and weighted averaging-partial least squares (WA-PLS) regression in pollen-based summer temperature transfer function calibration. We test the methods using a new, 113-sample calibration set from Estonia, Lithuania and European Russia, and a Holocene fossil pollen sequence from Lake Kharine, a previously studied lake in northeast European Russia. We find WA-PLS to outperform WA, probably because of smaller edge-effect biases in the ends of the calibration set gradient. The Bayesian-based calibration models show further improved performance compared with WA-PLS in leave-one-out cross-validation, while additional h-block cross-validation shows the Bayesian method to be little affected by spatial autocorrelation. Comparison with independent climate proxies reveals, however, some clear biases in the Bayesian palaeotemperature reconstructions, likely reflecting in part some specific limitations of our calibration set. As the selected prior parameters can significantly affect both Bayesian cross-validation performance and reconstructions, there is a clear need to further test the Bayesian method in different geographic contexts and over different timescales, with special attention given to the selection of the most realistic priors in each situation. In general, our finding that statistically well-performing transfer functions may produce clearly differing palaeotemperature reconstructions urges caution in transfer function-based inferences. We additionally test a spatially restricted, 58-sample subset of the full 113-sample calibration set. We find some reduced biases with the smaller set, likely because of complex, partially bimodal responses of several taxa along the longer temperature gradient, ill-suited for calibration methods assuming unimodal responses to climate.

## Keywords

calibration set size, Estonia, Lithuania, pollen, Russia, summer temperature

## Introduction

Transfer function-based studies have yielded many key insights into Quaternary environmental change, but several important challenges remain. First, the palaeoclimate reconstructions need to be validated by comparison with independent records (Birks and Birks, 2003). Second, the consistency of the modern calibration data, including the modern climate data, needs to be improved and tested (Daly, 2006; Kriticos and Leriche, 2010; Peterson and Nakazawa, 2008). Third, the various transfer function calibration methods need to be analysed and compared to assess their strengths and weaknesses in different situations, such as specific spatial or temporal scales (e.g. Heinrichs et al., 2001; Köster et al., 2004; Lotter et al., 2000; Paterson et al., 2002; Peyron et al., 2011). Fourth, the performance of transfer functions over longer timescales, with palaeoclimates and biotic assemblages less analogous with modern calibration data, needs to be critically evaluated (Jackson et al., 2009). Fifth, it is increasingly clear that more rigorous climatological, biogeographical and ecological consideration is needed for identifying the climatic parameters that can be realistically reconstructed in each region, from each particular biological proxy group, and from each climate-proxy calibration model, taking into account the possible effects of spatial autocorrelation when evaluating calibration models (Birks et al., 2010).

Weighted averaging-partial least squares regression (WA-PLS; ter Braak and Juggins, 1993) has been a commonly used transfer function calibration method in recent years, probably because WA-PLS generally (but not always) shows improved performance compared with simple two-way weighted-averaging regression (WA), another popular calibration method (Birks et al., 2010). A more recent alternative is offered by palaeoclimate reconstruction using Bayesian modelling. This approach is based on modelling the joint probability of the environmental variables of interest, the proxy data, and all the model parameters. A key component is

<sup>1</sup>University of Helsinki, Finland

<sup>2</sup>University of Oulu, Finland

<sup>3</sup>University of Bergen, Norway

<sup>4</sup>Nature Research Centre, Lithuania

<sup>5</sup>Herzen University, Russia

Received 20 April 2011; revised manuscript accepted 1 July 2011

## Corresponding author:

J Sakari Salonen, Department of Geosciences and Geography, PO Box 64,  
00014 University of Helsinki, Finland.  
Email: sakari.salonen@helsinki.fi

a 'forward' model that describes the data-generating process, that is, the causal effect of the environment on the climate proxy considered. Using the joint model, statistical inversion allows 'backward' inference that produces an environmental reconstruction consistent with the proxy data, together with its associated uncertainty. The first papers to use detailed Bayesian modelling for palaeoclimate reconstruction are Vasko et al. (2000), Toivonen et al. (2001), and Korhola et al. (2002). More recently Haslett et al. (2006), Erästö and Holmström (2006), and Holden et al. (2008) have also used a Bayesian approach. Other Bayesian approaches are discussed in Guiot et al. (2000) and Gachet et al. (2003). Bayesian techniques are also widely applied in many other areas that from the point of view of statistical modelling are similar to environmental reconstruction (e.g., Banerjee et al., 2004).

In pollen-based transfer function studies, a major division can be observed in terms of the spatial extent of the employed calibration data sets. While some studies use large-scale, continental or hemispheric data bases (e.g. Bordon et al., 2009; Cheddadi et al., 1998; Feurdean et al., 2008; Köhl et al., 2010; Sawada et al., 2004) others employ smaller, regional calibration sets (e.g. Finsinger et al., 2007; Goring et al., 2009; Seppä et al., 2004; St Jacques et al., 2008; Xu et al., 2010). While the first approach generally supports the use of the modern-analogue technique (MAT), a general objective of the latter approach is to try and select a regional calibration set with an optimal spatial size. Criteria involved in determining the desired calibration set pattern include the climate parameters which will be reconstructed and the spatial orientation of their gradients, the climatic range expected in the fossil data, and possible limitations of the calibration method used.

In this paper, we compare WA, WA-PLS and Bayesian modelling in pollen-summer temperature transfer function development, using a new 113-sample calibration set from European Russia, Estonia and Lithuania. The WA, WA-PLS and Bayesian transfer functions are compared by reconstructing summer temperature based on a Holocene fossil pollen sequence from Arctic European Russia. We focus on two main problems: First, the overall performance of Bayesian modelling versus WA and WA-PLS regression in transfer function calibration, and the relative vulnerability of these methods to spatial autocorrelation effects, with the presented calibration set; second, the effect of calibration set size (Bjune et al., 2010; Velle et al., 2011) on WA, WA-PLS and Bayesian transfer functions and the performance of these methods at different spatial scales.

## Study area

Our study area (Figure 1) covers a large part of northeast Europe, stretching from the Baltic Sea in the west to the Ural Mountains in the east, and from northwestern limit of the Eurasian Steppe in the south to the Barents Sea coast in the north. Most of the study area falls within the East European Plain with elevations mostly below 300 m a.s.l., with significantly higher elevations only found in the Urals. Four major vegetation zones cover the study area in latitudinal bands (Figure 1). In the north, a tundra zone stretches from the Barents Sea coast to the arctic treeline (major species: *Picea abies* ssp. *obovata*), located near the Arctic Circle. South on the treeline, a forest-tundra mosaic zone averaging c. 100 km in width is often distinguished before the transition to the boreal forest ('taiga'; major species: spruce (*Picea abies*), pine (*Pinus sylvestris*), birch (*Betula pendula* and *B. pubescens*) and larch (*Larix sibirica*)). Transition to mixed temperate forest (birch, alder (*Alnus incana*, *A. glutinosa*), lime (*Tilia cordata*), oak (*Quercus robur*) hazel (*Corylus avellana*) and maple (*Acer platanoides*)) occurs at c. 57–60°N. South of the mixed forest the steppe reaches up to 55°N in parts of European Russia, showing strong human influence with extensive crop and hay cultivation, and with scattered and often planted deciduous

forest stands. Mean annual temperature ranges from +7°C in Lithuania to –8°C in the Arctic Urals. Another major climatic gradient in the area is in the continentality index (Gorczyński, 1920, 1922) and the seasonal variation of temperature which increases considerably towards the east, driven by colder winters in the interior of the Eurasian landmass. Mean annual precipitation is relatively uniform, generally 600–700 mm, declining somewhat towards both the northern tundra and the southern steppe (Figure 1; Hijmans et al., 2005).

## Materials

### Surface pollen samples

The pollen–climate calibration data includes 113 lakes from Estonia, Lithuania and Russia. The 24 samples from Estonia were presented in Seppä et al. (2004), the rest of the samples are described for the first time here. Most sites (72) are located in the mixed forest zone, with 7 sites in the southern semi-steppe, 11 in the taiga, 7 in the forest–tundra transition zone, and 16 in the tundra north of the treeline (Figure 1). Samples were collected from medium-sized lakes in 1998–2009, using a surface sediment sampler from a boat or through the ice. Pollen samples were prepared using standard methods (Fægri and Iversen, 1989) with at least 500 terrestrial pollen and spores counted from each sample. Pollen nomenclature follows Moore et al. (1991).

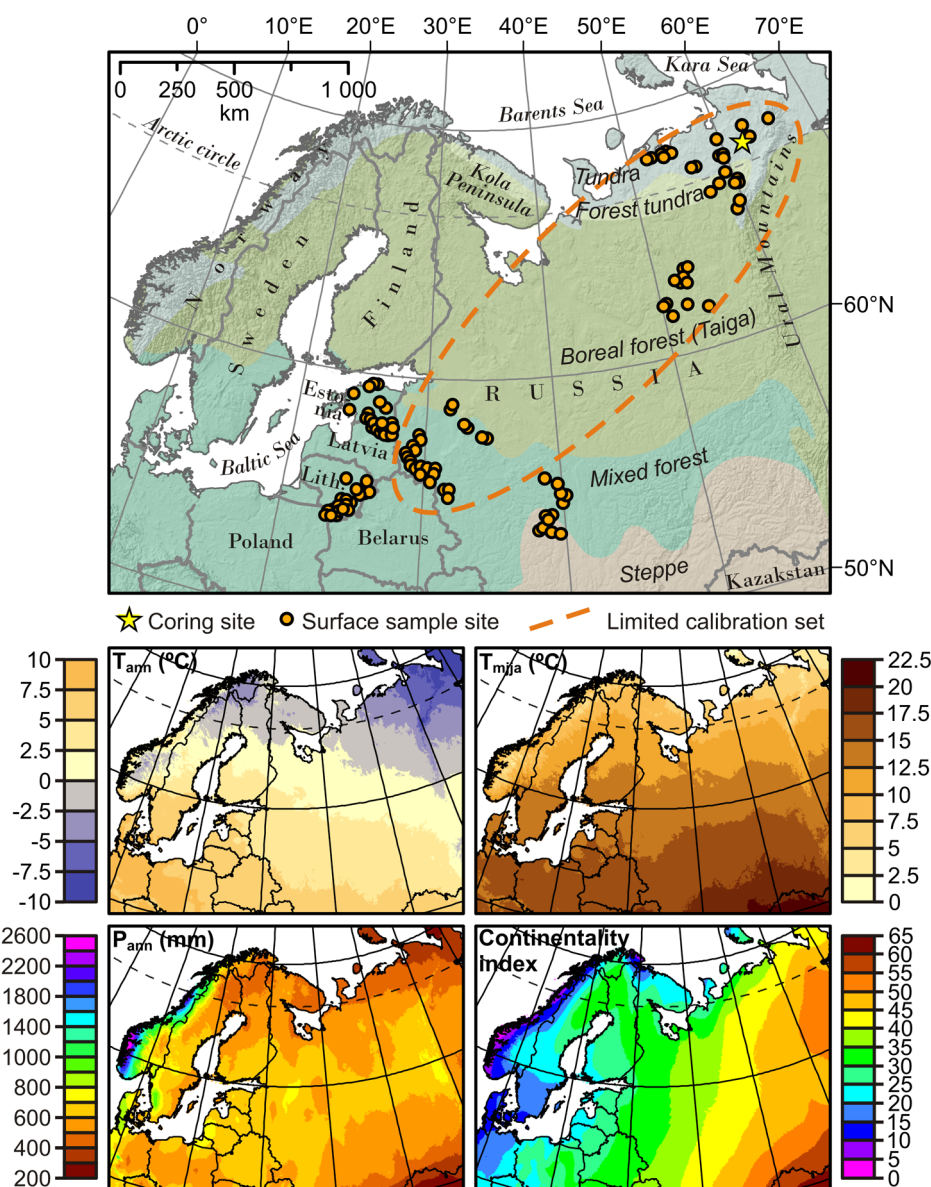
We constructed the transfer functions using two different calibration sets. The first calibration set (113S) uses all 113 surface samples. Figure 2 shows the variation of the most abundantly occurring taxa in 113S. The second calibration set (58S) is a spatially restricted 58-sample subset (indicated in Figure 1) of 113S. 58S omits samples in the southwestern end of 113S, included all steppe samples and a majority of the mixed-forest samples. The objective of testing the restricted subset is to simplify some of the complex or bimodal taxon responses to  $T_{mija}$  seen in 113S (Figure 2), and to test the effect of calibration set size on calibration model performance. A total of 104 pollen or spore types were identified in 113S, falling to 72 in 58S. All terrestrial pollen and spore types were included in the calibration sets.

### Modern climate data

The modern summer mean temperature (May-to-August mean =  $T_{mija}$ ) was determined for each calibration set sample.  $T_{mija}$  was chosen as the climate parameter as it roughly represents the temperature conditions during the growing season, including late spring, generally the most important single climate factor influencing the distribution, reproduction and growth of the plants in the arctic region (Euskirchen et al., 2009; Hinzman et al., 2004). The  $T_{mija}$  value for each surface sample site was extracted from a  $T_{mija}$  temperature grid (Figure 1), calculated in ArcInfo GIS software as the mean of WorldClim (Hijmans et al., 2005) mean temperature grids for May, June, July and August.  $T_{mija}$  values for the surface sites range from 17.3°C to 5.9°C, decreasing with a steady gradient towards the north (Figure 1).

### Fossil pollen data

The performance of our pollen–climate calibration set and the WA, WA-PLS and Bayesian modelling based transfer functions are tested on a Holocene pollen-stratigraphical record from Lake Kharinei (67°22'N 62°45'E), located near the Arctic Ural Mountains in European Russia (Figure 1). The general features of this sediment core, including its dating, and the ecological and palaeoclimatological context and implications are discussed fully in other papers (Jones et al., 2011; Salonen et al., 2011; Välranta et al., 2011). The Lake Kharinei fossil pollen sequence (Figure 3) includes 57 taxa of which 53 are found in 113S and 47 in 58S.



**Figure 1.** Map of the study area showing the location the calibration set surface samples and the Lake Kharinei coring site. The 58-sample subset of the full 113-sample calibration set is encircled. Vegetation types are synthesized from Khotinsky (1984), Rekeawicz (1998) and Olson et al. (2001). Climate parameters (from top left: mean annual temperature, May-to-August mean temperature, mean annual precipitation and the continentality index (Gorczynski, 1920, 1922)) are calculated based on WorldClim modern climate grids (Hijmans et al., 2005)

## Methods

### WA and WA-PLS transfer function calibration

Pollen–climate transfer functions (cf. Seppä and Birks, 2001; Seppä et al., 2004) were calculated using two-way weighted averaging (WA) with inverse deshrinking, as well as 2-, 3-, and 4-component weighted averaging-partial least squares (WA-PLS) regression (Birks, 1998; ter Braak and Juggins, 1993), separately for 58S and 113S. Calibration set species data values were square-root transformed for WA and WA-PLS regression to reduce noise in the data (Prentice, 1980). Calculation of WA and WA-PLS transfer functions was performed in the C2 computer programme (Juggins, 2007). The performance of all transfer functions was evaluated by leave-one-out cross-validation (Birks et al., 1990), and performance statistics were computed for each transfer function. These include the coefficient of determination ( $r^2$ ), root-mean-square error of prediction (RMSEP) and maximum bias.

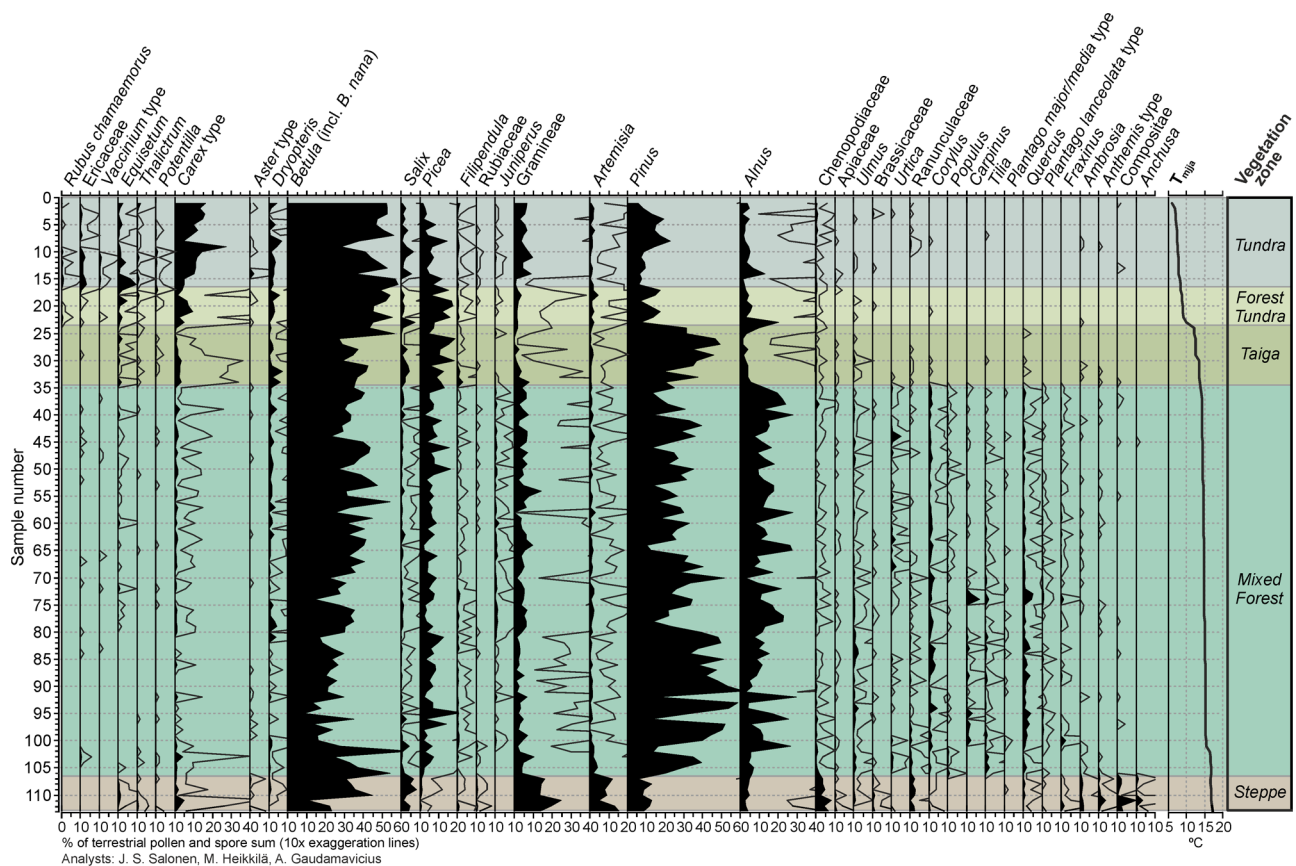
### Bayesian transfer function calibration

The Bayesian reconstruction method used here is Bummer, a Bayesian hierarchical multinomial regression model based on the classical

(direct) approach to calibration (Vasko et al., 2000). We describe here only the main idea; for further details and additional discussion we refer to the original paper, the application described in Korhola et al. (2002), as well as to the related work discussed in Erästö and Holmström (2006) and Haslett et al. (2006).

Suppose that the modern data are obtained from  $n$  calibration lakes and that the core data consist of  $N$  past pollen assemblages. We assume that from all sediment samples  $m$  different taxa are counted. Let  $x = (x_1, \dots, x_n)$  denote the modern temperatures and let the  $m \times n$  matrix  $Y$  contain the associated pollen taxon abundances. Similarly, the  $m \times N$  matrix of the past taxon abundances and the corresponding sequence of unknown past temperatures is denoted by  $Y^*$  and  $x^* = (x_1^*, \dots, x_N^*)$ , respectively.

The reconstruction of the past temperature time series  $x^*$  is described by its conditional probability density  $f(x^* | Y^*, Y, x)$ , given the observed counts  $Y^*, Y$  and the modern temperatures  $x$ . This is the so-called posterior density of  $x^*$  and it can be shown to be proportional to the integral  $\int f(x^*, \theta) f(Y^*, Y | x^*, \theta) d\theta$ . Here the quantity  $f(x^*, \theta)$  is the prior probability density of  $x^*$  and all the model parameters  $\theta$  and it is chosen by the modeller to reflect uncertainty about their values, prior to considering the data. The density  $f(Y^*, Y | x^*, \theta)$  is



**Figure 2.** Pollen diagram of the calibration set sites. Samples are ordered from top to bottom according to increasing  $T_{mija}$ . Coloured zones indicate the vegetation belts shown in the Figure 1 map. Taxa are ordered from left to right according to increasing  $T_{mija}$  optimum value in WA

the likelihood of the abundance data, given the temperatures and the parameters. It represents the forward causal part of the model. The backward step of Bayesian inference is represented by the computation of the posterior density of  $x^*$ . In practice, the particular model used is too complex to allow an analytical solution and inference is based on large sample from the posterior density obtained using Monte Carlo methods. For a general introduction to Bayesian modeling, see Gelman et al. (2004). A thorough account of Monte Carlo methods for Bayesian analysis is provided by Robert and Casella (2004).

The forward model assumes that for every pollen and spore type there is a temperature at which the corresponding host plant fares particularly well and that the pollen and spore relative abundances decay roughly symmetrically around this optimum temperature. In the Bummer model, the pollen specific response curve for taxon  $k$  is assumed to follow a unimodal Gaussian shape controlled by three parameters. This model assumption is coded in the quantities  $\lambda_{ik}$  defined as:

$$\lambda_{ik} = \alpha_k \exp \left[ - \left( \frac{\beta_k - x_i}{\gamma_k} \right)^2 \right] \quad (1)$$

where  $\alpha_k$  is a scaling factor,  $\beta_k$  the taxon-specific optimum temperature and  $\gamma_k$  the tolerance that represents sensitivity of the dependence of taxon abundance on the temperature. The relative magnitudes of the quantities  $\lambda_{ik}$  are used to specify the probabilities  $p_{ik}$  that random pollen grain or spore from site  $i$  belongs to taxon  $k$ . These probabilities are treated as random variables, which leaves room for variability between sites. Specifically, one assumes that the probabilities follow Dirichlet distribution,

$$(p_{i1}, p_{i2}, \dots, p_{im} | x_i, \{\alpha_k, \beta_k, \gamma_k\}) \sim \text{Dirichlet}(\lambda_{i1}, \dots, \lambda_{im}) \quad (2)$$

given the parameters  $\lambda_{i1}, \dots, \lambda_{im}$ .

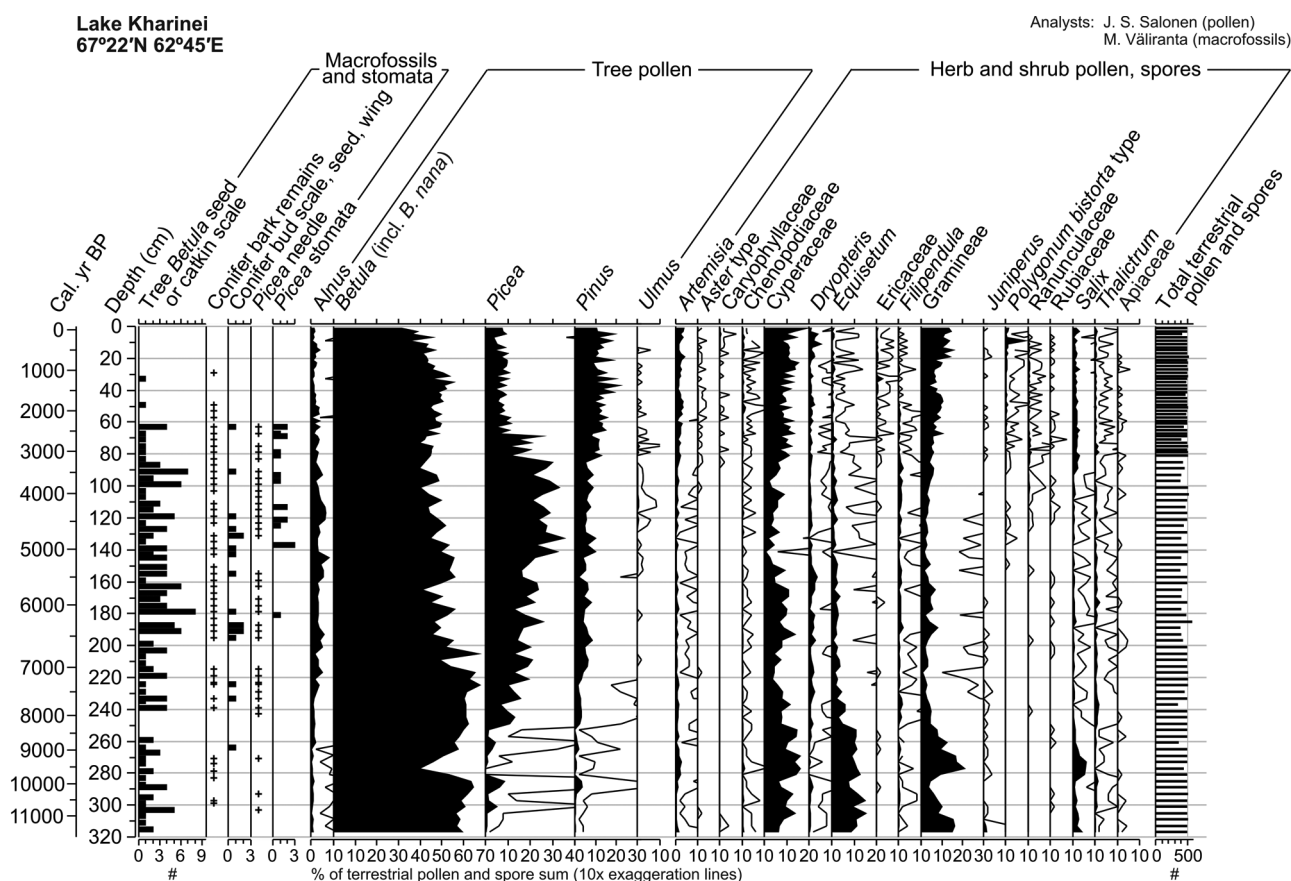
The causal forward model for the taxon abundances at site  $i$  is a multinomial distribution with the probabilities  $p_{ik}$ ,  $k = 1, \dots, m$ . The corresponding models involving the core relative abundances  $Y^*$ , the past temperatures  $x_i^*$  as well as the parameters  $\lambda_{ik}^*$  and  $p_{ik}^*$  are defined similarly. However, one assumes that the taxon specific parameters  $\alpha_k$ ,  $\beta_k$  and  $\gamma_k$  are the same for both modern and past abundances.

Vasko et al. (2000) and Korhola et al. (2002) both use the same gamma and uniform prior distributions for the tolerances  $\gamma_k$  and the scaling factors  $\alpha_k$ , respectively, and we make the same choices here:

$$\begin{aligned} \alpha_k &\sim \text{Unif}(0.1, 50), \\ \gamma_k &\sim \text{Gamma}(9, 3). \end{aligned} \quad (3)$$

For the past temperatures  $x_i^*$  and the optimum temperatures  $\beta_k$  Vasko et al (2000) and Korhola et al (2002) use normal priors centered on the modern 30-year mean July temperature. However, considering our long calibration set temperature gradient and the location of Lake Kharinei at its northern end, such a prior for  $\beta_k$  may be questionable, both in cross-validation and in reconstruction. Therefore, while we also report results based on the original Bummer-settings of Korhola et al. (2002), the actual model we use in our comparisons is based on species-specific priors for the optimum temperatures  $\beta_k$ . For reconstruction, using the idea underlying the WA method (ter Braak and Juggins, 1993), we estimated the optimal temperature for species  $k$  by  $\hat{\beta}_k = \sum_{i=1}^n \frac{y_{ik} x_i}{y_{+k}}$ , where  $x_i$  is the temperature at surface sample site  $i$ ,  $y_{ik}$  is the abundance for species  $k$  at site  $i$ , and  $y_{+k}$  denotes the total abundance for the species  $k$ . The prior distribution of  $\beta_k$  is then defined as

$$\beta_k \sim N(\hat{\beta}_k, (1.5\sqrt{3})^2) \quad (4)$$



**Figure 3.** Lake Kharinei biostratigraphy. Most abundantly occurring taxa are shown. Pollen and spore abundances are shown with black silhouette curves. Macrofossils counted in numbers (such as seeds) and conifer stomata are indicated by bars, whereas relative abundance of vegetative remains is indicated by a plus sign

In leave-one-out cross-validation the prior was constructed similarly, except that one site in turn is left out from the calibration set.

In cross-validation, choice of the prior for  $x_i^*$  is also somewhat problematic. The idea of cross-validation is to evaluate objectively the performance of a model that would be used in an actual reconstruction. The cross-validated models and the model used for reconstruction should therefore be essentially the same. Using the original Bummer approach, one is then led to set the prior means for the predicted temperatures equal to their actual values. Of course, if the prior is too tight, this may lead to over-optimistic results. On the other hand, making the temperature prior very uninformative, say uniform on the calibration set temperature gradient, does not make sense either because we would never make an actual reconstruction with such a model; after all, we should be allowed to use the fact that the Holocene temperatures probably do not deviate drastically from the current temperature. So therein lies the problem: the model that we actually use in reconstruction may not be reliably tested in cross-validation and the model that is perhaps more acceptable for cross-validation is not the one we would like to use in reconstruction. Consequently, as the basic idea behind true cross-validation is now impossible to implement, we find it best to report performance results using several prior settings that reflect different degrees to which information about the modern temperatures is utilized in the predictions. If  $i$  denotes the lake left out in leave-one-out cross-validation, the alternative prior settings tested are:

$$\text{Bayes 1: } \beta_k \sim N(T_{\text{mija}}(i), (\sqrt{3})^2), x_i^* \sim N(T_{\text{mija}}(i), 1) \quad (5)$$

$$\text{Bayes 2: } \beta_k \sim N(T_{\text{mija}}(i), (1.5\sqrt{3})^2), x_i^* \sim N(T_{\text{mija}}(i), 1.5^2) \quad (6)$$

$$\text{Bayes 3: } \beta_k \sim N(\hat{\beta}_k, (1.5\sqrt{3})^2), x_i^* \sim N(T_{\text{mija}}(i), 1.5^2) \quad (7)$$

$$\text{Bayes 4: } \beta_k \sim N(\hat{\beta}_k, (1.5\sqrt{3})^2), x_i^* \sim \text{Unif}(5.9, 15.6)(58S) \quad (8)$$

$$\text{or } x_i^* \sim \text{Unif}(5.9, 17.3)(113S)$$

Here  $T_{\text{mija}}(i)$  is the observed modern May-to-August mean temperature at lake  $i$ . The Bayes 1 model follows the original Bummer settings of Korhola et al. (2002) and Bayes 2 relaxes the prior of both the optimal temperatures  $\beta_k$  and the predicted temperatures by 50%. Both Bayes 3 and Bayes 4 use the species-specific priors for the optimum temperatures  $\beta_k$  but, where Bayes 3 adopts for  $x_i^*$  a normal prior centered on the predicted temperature, Bayes 4 employs an uninformative uniform prior on the calibration set temperature gradient.

While we report cross-validation results on all these four models, Bayes 3 seems like a reasonable compromise model whose cross-validation result probably is not overly optimistic but that could still be used in reconstruction. As noted above, the priors of  $\beta_k$  in the Bayes 1 and 2 models may be questionable for the calibration sets considered and the uniform priors of  $x_i^*$  in Bayes 4 almost completely disregard the information about the modern temperature. The Bayes 3 model can be interpreted to include *a priori* at least a range of  $\pm 3^\circ$  around the current temperature and, as this is believed to well cover the range of Holocene temperature variation for the sites considered; this prior choice is unlikely to bias cross-validation results much. Our results with the Bayes 4 model lend some support to this assumption.

Leave-one-out cross-validation was used to calculate  $r^2$ , RMSEP and maximum bias for the Bayesian transfer function. In the model description outlined above, the 'core' in this cross-validation setting can be thought of consisting of just one sediment layer ( $N=1$ ) with abundance data corresponding to the particular lake left out.

Predicting the temperature of just one lake took about 3.6 h of CPU time on Intel Core DuoE6750 computer, leading to a 17-day

cross-validation run for 113S. In Monte Carlo simulation, for each lake  $i$  we generated a sample of 20 000 realizations from the posterior distribution of  $x_i^*$  using the Metropolis-within-Gibbs algorithm (e.g. Robert and Casella, 2004). From this sample the first 5000 were used for algorithm burn-in and discarded, and from the rest every 10th sample was kept for the final analysis. The burn-in appeared to be more than sufficient for the algorithm to converge.

All computations for the Bayesian model were performed using the Matlab programming environment except that the highest posterior density intervals needed for the Bayesian error bars were calculated using the function `emp.hpd` from the R statistical software (R Development Core Team, 2009) package `TeachingDemos` (Snow, 2008).

### Spatial autocorrelation tests

The effect of spatial autocorrelation in the performance statistics measured in leave-one-out cross-validation (Telford and Birks, 2005, 2009) was evaluated by additionally performing h-block cross-validation (Telford and Birks, 2009) for each calibration method using 113S. In h-block cross-validation,  $T_{mjia}$  is estimated for each calibration set site using a calibration set which leaves out the site being predicted for, as well as all sites within the h-block radius  $r$ . The decline of  $r^2$  and the increase of RMSEP in h-block cross-validation relative to leave-one-out cross-validation figures is then used as a measure of the role of spatial autocorrelation in biasing leave-one-out cross-validation performance. For a h-block radius we chose  $r=100$  km, as a compromise that removes a moderate number of neighbouring sites, but leaves some analogues from each vegetation type in the calibration set. An average of 4.85 sites (min=0, max=15) were omitted in the h-block runs, with site density generally highest in the mixed-forest zone. The Bayesian model tested for spatial autocorrelation was Bayes 3 and we used for the taxon-specific parameters  $\alpha_k$ , and  $\gamma_k$  the prior distributions defined above (cf. the section on Bayesian transfer function calibration).

### Palaeoclimate reconstruction

The statistical significance of the  $T_{mjia}$  reconstruction was tested using the methods developed by Telford and Birks (2011a). We use redundancy analysis to determine if the proportion of the variance of the fossil data explained by the 113S-based WA and WA-PLS  $T_{mjia}$  reconstructions was larger than the proportion explained by most of 999 reconstructions of random environmental data. However, the WA- and WA-PLS-based tests are prone to Type I (false negative) error when the effective number of species is low (Telford and Birks, 2011a) which is the case with our pollen data (Hill's  $N2=3.45$ ). We therefore also test a MAT-based reconstruction (using five closest analogues, with squared chord distance as dissimilarity measure), as suggested by Telford and Birks (2011a). Significance test calculations were done using the R statistical software (R Development Core Team, 2009) package `palaeoSig` (Telford, 2011). Statistical significance of the Bayesian reconstruction was not tested due to the prohibitive computing demands of calculating separate models based on 999 sets of random environmental data.

With WA and WA-PLS, the reconstructed  $T_{mjia}$  values and their sample-specific standard errors were estimated using a 1000-cycle bootstrapping procedure (Birks, 2003) in the C2 software (Juggins, 2007). In Bayesian reconstruction we used for  $\alpha_k$ , and  $\gamma_k$  the same priors as in cross validations and

$$\beta_k \sim N(\hat{\beta}_k, (1.5\sqrt{3})^2), x_i^* \sim N(7.2, 1.5^2), \quad (9)$$

where  $7.2^\circ\text{C}$  is the observed modern mean for Lake Kharine. Thus, the Bayesian reconstruction transfer function corresponds to the Bayes 3 model tested in cross-validation. We generated a sample of 40 000 posterior realizations from the temperature vector  $x^*$  using

again Metropolis-within-Gibbs sampling. From those samples, the first 20 000 were used for burn-in and from the rest every 10th sample was kept for analysis. For each fossil sample we calculated the pointwise posterior expectation, and the pointwise 95% highest posterior density interval which can be thought to correspond to a classical confidence interval of  $\pm 2$  standard deviations.

## Results

We selected six calibration models for closer evaluation. First, we compare the WA model, the best-performing WA-PLS model, and the Bayes 3-based Bayesian model, all calibrated using 113S. Second, to analyse the effect of calibration set spatial extent on each method, we also test the WA, WA-PLS and the Bayesian models based on 58S. The models are evaluated based on their cross-validation performance and based on palaeotemperature reconstructions prepared with each model.

### Transfer function performance

Leave-one-out cross-validation performance statistics ( $r^2$ , RMSEP and maximum bias) for the developed transfer functions are shown in Table 1. WA-PLS shows improved performance over WA in all statistics. We use the van der Voet (1994)  $t$ -test to select the appropriate WA-PLS models for 58S and 113S, with WA-PLS components added as long as they provide a statistically significant improvement in predictive performance. With 113S the three-component WA-PLS model is selected and with 58S the two-component model is chosen. As expected, a tighter prior around the predicted temperature generally improves the cross-validation performance of the Bayesian models. When measured with  $r^2$  and RMSEP, the Bayesian transfer functions appear to be better than the non-Bayesian approaches, the only exception being a two-component WA-PLS model that marginally outperforms the uniform prior-based Bayesian transfer function on the smaller calibration set. However, as noted above, the results based on a uniform prior are too pessimistic, as one would not use such a model in practice. Except for the maximum bias, the effect of calibration set size on the Bayesian model performance is minor. The overall best model appears to be Bayes 3 which in the following will simply be referred to as the 'Bayesian model'.

Figure 4 shows the plots of predicted  $T_{mjia}$  versus the observed modern  $T_{mjia}$  and residuals of predicted  $T_{mjia}$  in leave-one-out cross-validation with the selected models. Generally, the Bayesian models show smaller scatter of temperatures within each vegetation type, compared with WA and WA-PLS. In tundra samples ( $T_{mjia} < 8^\circ\text{C}$ ) WA and WA-PLS models show their worst performance with the 113S WA model overestimating temperatures by up to  $4^\circ\text{C}$  and the 113S WA-PLS model by up to  $3^\circ\text{C}$ , while the residuals are also mostly positive but somewhat smaller with 58S. The Bayesian models perform significantly better in the tundra. In taiga samples ( $T_{mjia} 10\text{--}14^\circ\text{C}$ ) WA-PLS and Bayesian models show widely scattered residuals between  $-3$  and  $+3^\circ\text{C}$  but without systematic bias, while the WA models tend to underestimate temperatures by up to  $3^\circ\text{C}$ . All models perform well in mixed forest ( $T_{mjia} 14\text{--}16.7^\circ\text{C}$ ). In the steppe samples ( $T_{mjia} > 16.7^\circ\text{C}$ ) only included in 113S, all models underestimate temperatures, with greater negative bias in WA compared with WA-PLS and Bayesian models.

The h-block cross-validation performance statistics for the 113S-based WA, WA-PLS and Bayesian models are shown in Table 2, with the respective leave-one-out cross-validation figures. The performance of WA and WA-PLS worsens moderately in h-block cross-validation, with some rise in RMSEP and maximum bias, and the decline of  $r^2$  by  $c. 0.1$  with both models. At least a part of this decline is due to the removal of environmentally similar, spatially close analogues. The Bayesian model is less affected compared with WA and WA-PLS, with an  $r^2$  drop of just 0.04 in h-block cross-validation.

**Table 1.** Performance statistics for WA, WA-PLS and Bayesian pollen–climate models for  $T_{mija}$  based on leave-one-out cross-validation. Reported statistics are root mean square error of prediction (RMSEP), coefficient of determination ( $r^2$ ), and maximum bias. Shown in italics are the WA, WA-PLS and Bayesian models selected for further analysis with each calibration set size

Model	113 Calibration sites			58 Calibration sites		
	RMSEP	$r^2$	Max. bias	RMSEP	$r^2$	Max. bias
WA	<i>1.270°C</i>	<i>0.819</i>	<i>2.993°C</i>	<i>1.356°C</i>	<i>0.827</i>	<i>2.388°C</i>
WA-PLS, 2 components	1.126°C	0.858	2.448°C	1.121°C	0.882	1.629°C
WA-PLS, 3 components	<i>1.043°C</i>	<i>0.878</i>	<i>1.367°C</i>	1.206°C	0.863	1.536°C
WA-PLS, 4 components	1.064°C	0.873	1.299°C	1.371°C	0.828	1.224°C
Bayes 1	0.660°C	0.954	1.577°C	0.738°C	0.956	1.992°C
Bayes 2	0.836°C	0.922	1.784°C	0.846°C	0.935	2.098°C
Bayes 3	<i>0.801°C</i>	<i>0.930</i>	<i>1.178°C</i>	<i>0.878°C</i>	<i>0.930</i>	<i>1.788°C</i>
Bayes 4	0.998°C	0.892	2.226°C	1.171°C	0.879	2.349°C

### Palaeotemperature reconstructions

The WA and WA-PLS  $T_{mija}$  reconstructions test as marginally non-significant (for WA,  $p = 0.069$ ; for WA-PLS,  $p = 0.076$ ). The MAT reconstruction, however, tests as significant ( $p = 0.004$ ). The main features of the MAT reconstruction (supplementary data, available online) closely resemble those of the WA and especially the WA-PLS reconstructions, suggesting that the non-significance of the WA and WA-PLS reconstructions may be due to Type I error (Telford and Birks, 2011a).

Lake Kharinei  $T_{mija}$  reconstructions based on the WA, WA-PLS and Bayesian models calibrated with 113S and 58S are presented in Figure 5. Both WA-PLS based reconstructions show a prominent Holocene thermal maximum (HTM) during the mid Holocene, at *c.* 8000–3000 cal. yr BP, with  $T_{mija}$  around 10°C, *c.* 3°C above the present-day value. The differences in the 113S and 58S WA-PLS reconstructions are minor, the major difference being that the 113S-based reconstruction shows slightly flattened curve due to higher reconstructed early- and late-Holocene temperatures. WA shows similar temperatures compared with WA-PLS during the mid Holocene, but during both the early and late Holocene temperatures are significantly higher with WA, leading to a much reduced Holocene temperature range (*c.* 1.5°C). As with WA-PLS, switching from 113S to 58S in WA decreases early- and late-Holocene temperatures slightly. The Bayesian reconstructions have some major differences compared with the WA-PLS reconstructions. They share with the WA-PLS reconstructions the major feature that the mid Holocene (*c.* 7000–3000 cal. yr BP) is indicated as the warmest period. However, the  $T_{mija}$  values during the HTM are significantly lower in the Bayesian reconstructions at *c.* 8.5–9°C, and the cooling indicated since HTM is *c.* 1°C in the Bayesian reconstructions compared with *c.* 3°C in the WA-PLS based reconstructions. The effect of calibration set size on the Bayesian reconstruction is negligible.

LOESS smoothers (span 0.25, one robustifying iteration) were fitted through the WA, WA-PLS and Bayesian based reconstructions. Smoothing is here viewed as an exploratory tool that can be used to reveal the salient features of reconstructions. Note, however, that since the result of a Bayesian reconstruction is in fact the full probability density of possible past temperatures that are consistent with the data, the posterior mean and its LOESS smooth represent only simple summary aspects of the full reconstruction. It can be argued that it is actually more meaningful to explore *several* different smooths of the *whole* posterior distribution since then one can make inferences about the salient features of past temperature variation and their statistical significance in many different time scales (cf. Erästä and Holmström, 2006).

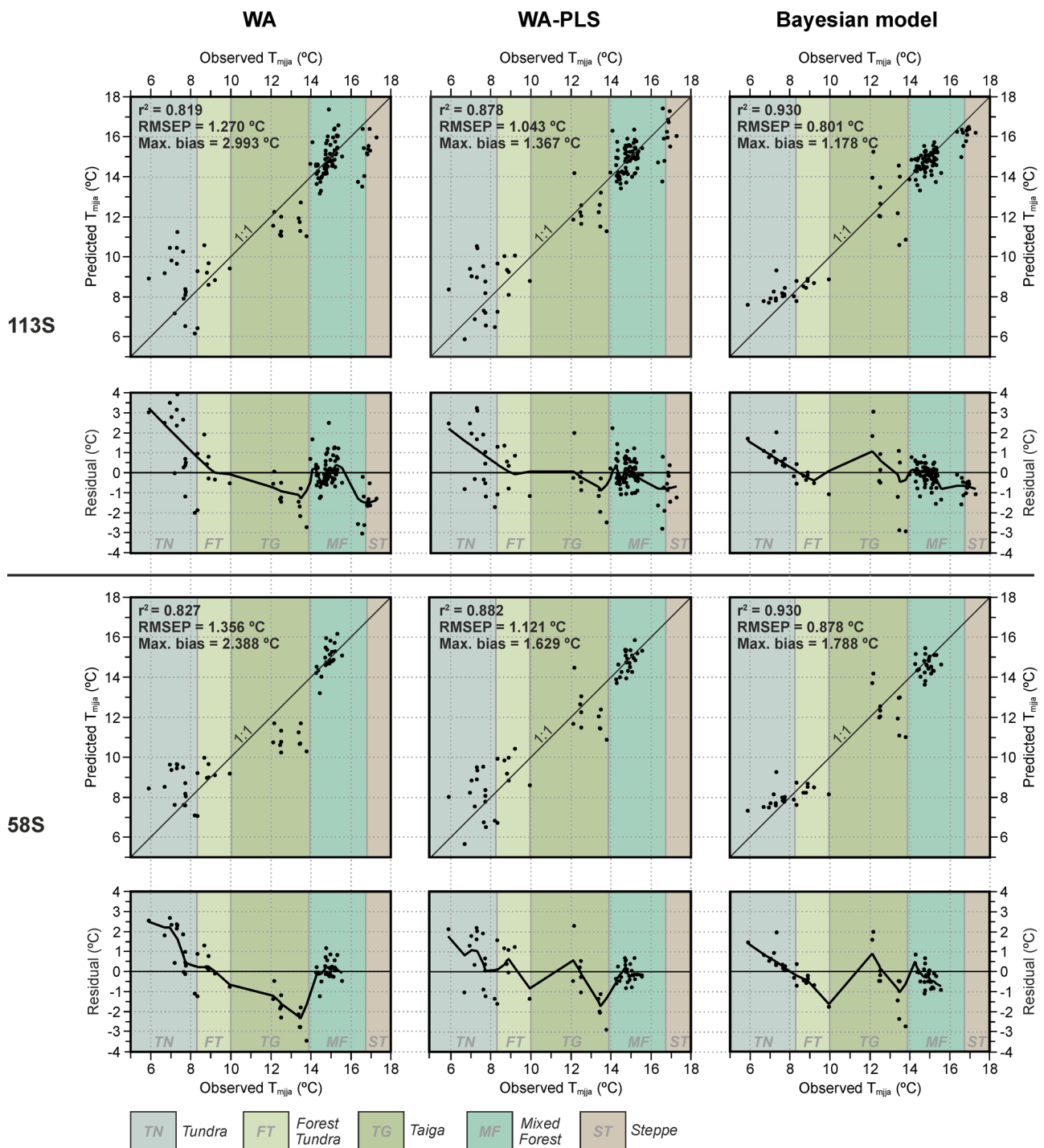
## Discussion

### The effect of calibration set size

Examination of the calibration set pollen curves (Figure 2) reveals several taxa with complex responses to  $T_{mija}$ . For example, some

non-arboreal pollen (NAP) taxa show responses to  $T_{mija}$  with a clear bimodal component. The most important of these taxa is Gramineae, which shows a clear rise in the north, moving from taiga forests into the tundra. However, the greatest values of Gramineae are found in the steppe, south of the mixed forest zone. Similar behaviour is shown by *Carex* type and *Equisetum* (Figure 2) with strong maxima in the tundra but also rising in the steppe. All of these are broad taxa that include species typical of widely different environments, resulting in such complex behavior when observed at family or genus level (Seppä et al., 2004). These complicated responses to the environmental variable being reconstructed present a potentially significant problem for transfer function calibration. A pollen curve such as that of Gramineae (Figure 2) intuitively has significant indicator value – in the north near the Arctic treeline the family clearly indicates tundra conditions instead of taiga, and in the south it is strongly associated with steppe as opposed to mixed forest. However WA and WA-PLS, as well as the Bayesian method model taxon responses unimodally, and are thus unable to capture this information. For example, WA calculates for Gramineae a single temperature optimum of 13.7°C associated with the southern taiga, an environment not typical of the taxon. Through such unrepresentative optima, bimodal taxon responses can cause biases in temperature estimates for pollen assemblages rich in these taxa. While WA-PLS is generally able to reduce biases compared with WA by adjusting the optima of the taxa abundant in high-residual samples (Birks, 2003; ter Braak and Juggins, 1993), this process still results in loss of information as bimodal real-world responses are modelled unimodally. For Bayesian modeling, we also examined the relationship between the prior and the posterior distributions of the optimal temperatures  $\beta_k$  and the tolerances  $\gamma_k$  for taxa that appeared only in few calibration lakes and fossil samples. Contrary to what one might expect, the prior did not dominate the posterior for such taxa showing that the prior distributions were sufficiently vague to allow even a relatively small number of data to have an effect on the posterior.

One way to mitigate the biases stemming from bimodal responses is to limit the spatial extent of the calibration set (cf. also Velle et al., 2011). Our smaller calibration set 58S omits the steppe samples and some of the mixed forest samples. The omission of steppe samples gives the aforementioned NAP taxa unimodal responses with maxima in the tundra. In both WA and WA-PLS, 58S gives improved performance over 113S in tundra temperatures in terms of smaller positive residuals, with a greater change towards smaller bias in WA than in WA-PLS. Significant positive residuals remain in the tundra in 58S, however, likely partially due to the so-called edge effects (cf. Birks, 2003; ter Braak and Juggins, 1993) which cause overestimation of optima for the coldest taxa and underestimation for the warmest taxa. With WA-PLS, overall performance is similar with 58S compared to 113S (Table 1) although a simpler model was used (two components versus three). The differences in the Bayesian models based on 113S and 58S are small, with a slight increase in

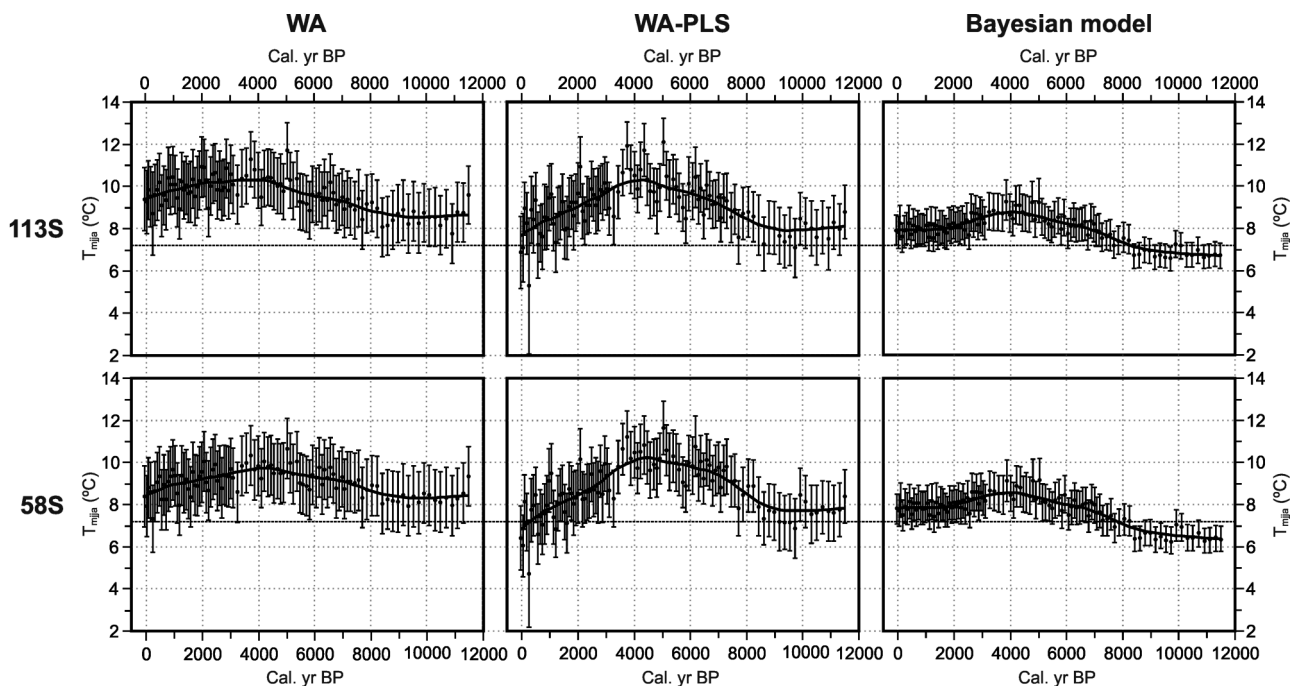


**Figure 4.** Plots of predicted versus observed  $T_{mja}$  and residuals of predicted versus observed  $T_{mja}$  for six transfer functions in leave-one-out cross-validation. Transfer functions are prepared with weighted averaging (WA) and weighted averaging-partial least squares (WA-PLS) regression and Bayesian modelling. Two different calibration sets are used with each calibration method: 113S uses all 113 surface samples, 58S is a 58-sample subset of 113S. Coloured zones indicate the vegetation belts shown in Figure 1. LOESS smoothers (span 0.25, one robustifying iteration) are added to the residual plots. Reported performance statistics for each transfer function are coefficient of determination ( $r^2$ ), root mean square error of prediction (RMSEP), and maximum bias

$r^2$  with 58S compared with 113S. In the WA- and WA-PLS-based Lake Kharine palaeotemperature reconstructions, the smaller tundra residuals with 58S lead to slightly lower temperatures before *c.* 10 000 cal. yr BP and after *c.* 2000 cal. yr BP (Figure 5), the periods with most NAP-rich fossil assemblages (Figure 3). The removal of the steppe samples in 58S improves the indicator value of the tundra taxa showing complex responses over the longer climatic gradient of 113S, leading to lower bias in the transfer function, and probably a more credible palaeotemperature reconstruction from fossil tundra assemblages.

#### Bayesian transfer functions

Bayesian modelling offers some notable advantages over more classical methods. First, ecological knowledge can be explicitly embedded in the models. Second, the structure of the model formulated in terms of probabilities is transparent and the use of probability distributions instead of point estimates provides more information about the quantities of interest, including a rigorous assessment of the uncertainties associated with them. Third, the interpretation of the results is straightforward in the Bayesian approach. For example, Bayesian error intervals have direct interpretation in terms of the



**Figure 5.** Lake Kharinei  $T_{mjia}$  reconstructions using weighted averaging (WA) and weighted averaging-partial least squares (WA-PLS) regression and Bayesian modelling. Reconstructions are prepared using two different calibration sets with each method: 113S uses all 113 surface samples, 58S is a 58-sample subset of 113S. Errors bars show the bootstrap-estimated standard errors for the WA and WA-PLS reconstructions, and the sample-specific 95 % error for the Bayesian reconstruction. LOESS smoothers (span 0.25, one robustifying iteration) are added to the reconstructions

posterior probability. The corresponding concept in non-Bayesian statistics is based on the idea of repeated experiments, a view that can have serious difficulties in ecological applications where the number of replications is small (Ellison, 1996; Vasko et al., 2000). In our experiments, the calibration set prediction performance of the Bayesian model appeared to be somewhat better than that of the WA-PLS reconstruction method tested. Possible explanations for the better performance include taking into consideration the uncertainty concerning site specific latent variables (the probabilities  $p_{ik}$  and  $p_{ik}^*$  in the model) and explicit utilization of ecological background knowledge (cf. Vasko et al., 2000).

In Bayesian modelling, one should always check the sensitivity of the results to the selection of particular priors for the parameters. Both in Korhola et al. (2002) and Eröstö and Holmström (2006) the prior distribution for the  $x_k$ s had the largest influence on the reconstructions but the choice of prior distributions for  $\alpha_k$ ,  $\beta_k$  and  $\gamma_k$  was not that important. In the cross-validation experiment, we tested four different models that in their prior setting employ the information about the modern temperatures  $T_{mjia}$  to varying degrees. Using a vague uniform temperature prior that covers the entire  $T_{mjia}$  gradient naturally degrades the cross-validation performance of the Bayesian model. Such a transfer function nevertheless was still competitive against the WA and the WA-PLS models. For closer evaluation we selected a Bayesian transfer function that can be seen as a compromise between the Bummer settings of Korhola et al. (2002) that, for the present calibration sets, might produce too optimistic results, and a vague prior model that would be undesirable in an actual reconstruction.

Lake Kharinei reconstructions were also made with various prior standard deviations for the past temperature prior, including the values 1.0°C and 1.5°C tried in cross-validation, as well as with different prior means. The results obtained were not very sensitive to prior selection.

Despite its great potential, Bayesian modelling has not been extensively used in palaeoclimate reconstructions. One reason for its modest popularity may have been the computing demands which can be substantial when large proxy data sets are considered. However, with rapidly increasing computing power offered by modern

PCs the situation has changed considerably during the past decade making the Bayesian approach more attractive. Given the total effort put into collecting and analysing the proxy data, it could be argued that if the Bayesian reconstruction model works well, it is worth the computation times needed. Still, better simulation techniques and suitable model approximations could be helpful in making the Bayesian approach more accessible (cf. Bhattacharya, 2004; Haslett et al., 2006).

#### Evaluation of Bayesian modelling versus WA and WA-PLS

While the WA-, WA-PLS- and Bayesian-based palaeotemperature reconstructions show broadly similar trends, there are major differences in the absolute reconstructed temperatures during different time periods. The major difference between the WA- and WA-PLS-based reconstructions is that WA indicates significantly higher temperatures from early- and late-Holocene tundra conditions compared with WA-PLS, whereas reconstructed temperatures from the mid-Holocene taiga assemblages are similar with both methods. These differences are likely explained by the large, positive cross-validation residuals of the WA-based models for tundra samples, leading to overestimation of temperatures from fossil tundra assemblages. While the WA-based reconstructions can thus be diagnosed as likely less reliable than those based on WA-PLS, it is striking how large the differences between the WA-PLS- and Bayesian-based reconstructions are, considering that both methods have excellent nominal performance in cross-validation, comparing favourably with pollen-based summer temperature ( $T_{jul}$ ) transfer functions prepared from other geographic regions (cf. Birks and Seppä, 2004; Birks et al., 2010; Seppä and Bennett, 2003). Clearly at least one (or possibly both) of the 'excellent' transfer functions is in fact giving heavily biased results.

Validation of palaeoclimate inferences is a major concern for quantitative palaeoclimatology (e.g. Barber and Langdon, 2007; Bennion et al., 1995; Birks and Birks, 2003; Telford and Birks, 2011a). A key method is to compare the reconstruction with another, independent climate proxy (Birks, 2003). For the Lake Kharinei record plant macrofossils provide one such way to validate the

**Table 2.** Comparison leave-one-out and h-block ( $h=100$  km) cross-validation performance statistics for selected WA-, WA-PLS- and Bayesian-based calibration models for  $T_{mija}$ , using all 113 calibration set sites. Reported statistics are root mean square error of prediction (RMSEP), coefficient of determination ( $r^2$ ), and maximum bias

	Leave-one-out			h-Block ( $r = 100$ km)		
Model	RMSEP	$r^2$	Max. bias	RMSEP	$r^2$	Max. bias
WA	1.270°C	0.819	2.993°C	1.593°C	0.717	3.119°C
WA-PLS, 3 components	1.043°C	0.878	1.367°C	1.392°C	0.786	1.991°C
Bayes 3	0.801°C	0.930	1.178°C	0.990°C	0.894	1.655°C

WA-PLS- and Bayesian-based reconstructions (cf. Birks and Birks, 2003). Based on plant macrofossils (Salonen et al., 2011) as well as pollen accumulation rates (Välranta et al., 2011), the northern limit of the taiga forest is suggested to have reached Lake Kharinei in the mid Holocene, c. 8000–2500 cal. yr BP. The northern limit of the taiga forest is today associated with a  $T_{mija}$  value of c. 9°C, suggesting that the WA-PLS-based HTM value may be realistic whereas the Bayesian-based HTM temperatures may be too low. In other studies from this part of Northern Russia, several lines of evidence have been invoked to suggest a cooling of 2–3°C or more from the HTM to the present day. These include the geographical shift of vegetation and permafrost zones (Kultti et al., 2003, 2004; MacDonald et al., 2000; Oksanen et al., 2001; Välranta et al., 2003), indicator species macrofossils (Kultti et al., 2004), and quantitative reconstructions based on pollen (Andreev and Klimanov, 2000; Andreev et al., 2005) and chironomids (Andreev et al., 2005). The slight cooling of c. 1°C in the Bayesian reconstructions is unlikely to account for these environmental changes, including a treeline withdrawal of 150 km or more (MacDonald et al., 2000; Salonen et al., 2011) in the Pechora Basin, whereas the cooling of 3°C in the WA-PLS reconstruction may be more realistic. We consider the amplitude of late-Holocene cooling in the Bayesian model to be too small, possibly explained by too low reconstructed HTM temperatures.

We also computed the posterior probabilities of the HTM (8000–3000 cal. yr BP) temperatures reconstructed by the WA and WA-PLS models. The results (supplementary data, available online) confirm that these temperatures indeed are significantly higher than the reconstructions produced by the Bayesian model. Assuming that the Bayesian model is correct, the posterior probability of observing temperatures at least as high as the WA temperatures is less than 5% for all fossil samples. For WA-PLS, only 5 out of 42 fossil samples produce reconstructions that could be exceeded by the Bayesian model with posterior probability higher than 5%.

Spatial autocorrelation (Legendre, 1993) in the smooth  $T_{mija}$  gradient of our study area likely accounts for some of the good performance of our calibration models (cf. Telford and Birks, 2005, 2009), especially in the densely spaced mixed-forest samples, and thus inflates the  $r^2$  and RMSEP figures to some degree. However, based on our tests none of the three methods (WA, WA-PLS and Bayesian modelling) is greatly affected by spatial autocorrelation with this calibration set (Table 2), with the Bayesian model in particular showing only a slight decrease in performance.

To shed light on the reasons for differences between the reconstructions obtained with WA-PLS and Bayesian modelling, we compared the pollen compositions in the core samples and in the calibration set. As a measure of analogue quality we used the squared chord distance (Overpeck et al., 1985) between the vectors of relative abundances of the fossil and calibration set data. As was to be expected on the basis of the northern location of Lake Kharinei (Figure 1), the closest analogues with the core samples were found among the calibration set lakes in the colder end of the temperature

gradient (see supplementary data, available online, for the list of analogues). Further exploration suggested that, for the Bayes 1 and Bayes 2 models that use priors similar to those in Vasko et al. (2000) and Korhola et al. (2002), the reconstruction relies rather heavily on the few closest analogues only, overlooking warmer lakes that still are relatively close in the Euclidean metric. As this may result in overfitting that produces too low reconstructed temperatures, we examined also the effects of prior choices of the parameters  $\alpha_k$ ,  $\beta_k$  and  $\gamma_k$ , and concluded that only the prior of  $\beta_k$  has a noticeable influence on the reconstruction. One motivation for using a different prior for  $\beta_k$  in the Bayes 3 and Bayes 4 models was to avoid this potential bias towards colder temperatures. Compared with the Bayes 2 setup, the reconstructed temperatures with species-specific  $\beta_k$  priors were on average approximately 0.25°C higher but the salient features of the temperature time series did not change.

Good figures for statistics like  $r^2$  and RMSEP may hide significant biases in specific environment types. In our calibration models, a majority of surface samples is from the mixed forest zone (Figures 1 and 2). The excellent prediction accuracy for the mixed forest samples weighs heavily in the calculated  $r^2$  and RMSEP, belying the performance in tundra and taiga which, while satisfactory, is significantly worse than in mixed forest (Figure 4; cf. Telford and Birks, 2011b). As the largest mismatch between WA-PLS- and Bayesian-based reconstructions is during the mid-Holocene taiga stage as opposed to the early- and late-Holocene forest-tundra/tundra stages, it is possible that the main discrepancies stem from underestimation of palaeotemperatures specifically from taiga assemblages by the Bayesian method. While the Bayesian models perform well along much of the climatic gradient, the Bayesian cross-validation residuals in taiga are among the largest seen in this study, especially with 113S (Figure 4). The Bayesian taiga residuals range from –3.0°C to +3.1°C, and while there is no systematic negative bias for the modern samples, it seems likely that temperatures from the specific fossil taiga assemblages encountered in the Lake Kharinei core are underestimated by the Bayesian method. The decrease of performance by the Bayesian transfer functions in taiga environments may reflect in part a specific shortcoming of our calibration set. The taiga samples are separated from the rest of the calibration set by large spatial and environmental gaps (Figure 1), which likely contributes to biases in calculated optima for taiga taxa and in reconstructed temperatures from fossil taiga assemblages. The overall performance of the Bayesian models is highly promising, and this reconstruction method should be further evaluated, including the assessment of the prior parameter selection principles, and including testing with new calibration sets in different geographic areas.

## Conclusions

- We developed pollen-based calibration models for summer temperature using WA and WA-PLS regression, as well as the little-used Bayesian modelling-based method. The prepared transfer functions were tested in palaeoclimate reconstruction based on a Holocene fossil pollen sequence from northeast European Russia. The Bayesian models show significantly improved performance in leave-one-out cross-validation over WA and WA-PLS, while based on h-block-cross-validation the Bayesian performance is only slightly affected by spatial autocorrelation. However, comparison of the reconstructions with several independent records reveals some clear biases in the Bayesian palaeotemperatures. The biases in the Bayesian reconstructions likely reflect in part some specific deficiencies of our calibration data, emphasizing the need to further test the Bayesian method using other calibration sets.
- As the priors can significantly influence both the Bayesian cross-validation performance and the final reconstruction, a key challenge in future Bayesian work is the identification of ecologically

and climatologically realistic prior parameters, in terms of the timescale and the geographic context of each reconstruction scenario.

- We found calibration models based on a geographically limited subset of the full calibration set to show some smaller biases in cross-validation. This is probably due to the simpler ecological response of several taxa on the smaller spatial scale, easier modelled on methods which assume unimodal taxon responses.
- While the WA-PLS and Bayesian reconstructions presented in this study show similar trends, there are major differences in the absolute reconstructed temperatures, suggesting that caution should be exercised with apparently well-performing transfer functions, as they may produce significantly biased palaeoclimate reconstructions, even when statistical performance has not been significantly inflated by spatial autocorrelation effects. This suggests that far-reaching inferences should be avoided when working with any single reconstruction method, and underlines the importance of validation of quantitative reconstructions by comparison with independent proxies.

## Acknowledgements

Vivienne J. Jones, Nikolay Letuka, Olga Malozemova, Vasily Ponomarev, and Angela Self are thanked for help during field-work, Maija Heikkilä for help with the pollen analysis, and Minna Väiliranta for providing the macrofossil data.

## Funding

This study was financed by the CARBO-North project (EU Sixth Framework Programme, Global Change and Ecosystems sub-programme, project number 036993), the Academy of Finland project 1132588 (Quantitative Vegetation Reconstructions), the Finnish Graduate School of Geology, and the Finnish Doctoral Programme in Stochastics and Statistics.

## References

- Andreev AA and Klimanov VA (2000) Quantitative Holocene climatic reconstruction from Arctic Russia. *Journal of Paleolimnology* 24: 81–91.
- Andreev AA, Tarasov PE, Ilyashuk BP, Ilyashuk EA, Cremer H, Hermichen W-D et al. (2005) Holocene environmental history recorded in Lake Lyadhej-To sediments, Polar Urals, Russia. *Palaeogeography, Palaeoclimatology, Palaeoecology* 223: 181–203.
- Banerjee S, Carlin B and Gelfand A (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall.
- Barber KE and Langdon PG (2007) What drives the peat-based palaeoclimate record? A critical test using multi-proxy climate records from northern Britain. *Quaternary Science Reviews* 26: 3318–3327.
- Bennion H, Wunsam S and Schmidt R (1995) The validation of diatom-phosphorus transfer functions: An example from Mondsee, Austria. *Freshwater Biology* 34: 271–283.
- Bhattacharya S (2004) Importance Resampling MCMC: A methodology for cross-validation in inverse problems and its applications in model assessment. PhD thesis, Trinity College Dublin.
- Birks HJB (1998) Numerical tools in quantitative palaeolimnology – Progress, potentialities, and problems. *Journal of Paleolimnology* 20: 301–332.
- Birks HJB (2003) Quantitative palaeoenvironmental reconstructions from Holocene biological data. In: Mackay A, Battarbee R, Birks J and Oldfield F (eds) *Global Change in the Holocene*. London: Arnold, 107–123.
- Birks HH and Birks HJB (2003) Reconstructing Holocene climates from pollen and plant macrofossils. In: Mackay A, Battarbee R, Birks J and Oldfield F (eds) *Global Change in the Holocene*. London: Arnold, 342–357.
- Birks HJB and Seppä H (2004) Pollen-based reconstructions of late-Quaternary climate in Europe – progress, problems and pitfalls. *Acta Palaeobotanica* 44: 317–334.
- Birks HJB, Heiri O, Seppä H and Björne AE (2010) Strengths and weaknesses of quantitative climate reconstructions based on late-Quaternary biological proxies. *The Open Ecology Journal* 3: 68–110.
- Birks HJB, Line JM, Juggins S, Stevenson AC and ter Braak CJF (1990) Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society B* 327: 263–278.
- Björne AE, Birks HJB, Peglar SM and Odland A (2010) Developing a modern pollen–climate calibration data set for Norway. *Boreas* 39: 674–688.
- Bordon A, Peyron O, Lézine A-M, Brewer S and Fouache E (2009) Pollen-inferred Late-Glacial and Holocene climate in southern Balkans (Lake Maliq). *Quaternary International* 200: 19–30.
- Cheddadi R, Kamakowa K, Guiot J, de Beaulieu J-L, Reille M, Andrieu V, Granaszewski W et al. (1998) Was the climate of the Eemian stable? A quantitative climate reconstruction from seven European pollen records. *Palaeogeography, Palaeoclimatology, Palaeoecology* 143: 73–85.
- Daly C (2006) Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology* 26: 707–721.
- Ellison AM (1996) An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6: 1036–1046.
- Erästä P and Holmström L (2006) Selection of prior distributions and multiscale analysis in Bayesian temperature reconstructions based on fossil assemblages. *Journal of Paleolimnology* 36: 69–80.
- Euskirchen ES, McGuire AD, Chapin FS III, Yi S and Thompson CC (2009) Changes in vegetation in northern Alaska under scenarios of climate change, 2003–2100: Implications for climate feedbacks. *Ecological Applications* 19: 1022–1043.
- Fægri K and Iversen J (1989) *Textbook of Pollen Analysis*. Chichester: Wiley.
- Feurdean A, Klotz S, Mosbrugger V and Wohlfarth B (2008) Pollen-based quantitative reconstructions of Holocene climate variability in NW Romania. *Palaeogeography, Palaeoclimatology, Palaeoecology* 260: 494–504.
- Finsinger W, Heiri O, Valsecchi V, Tinner W and Lotter AF (2007) Modern pollen assemblages as climate indicators in southern Europe. *Global Ecology and Biogeography* 16: 567–582.
- Gachet S, Brewer SS, Cheddadi R, Davis B, Gritti E and Guiot J (2003) A probabilistic approach to the use of pollen indicators for plant attributes and biomes: An application to European vegetation at 0 and 6 ka. *Global Ecology and Biogeography* 12: 103–118.
- Gelman A, Carlin JB, Stern HS and Rubin DB (2004) *Bayesian Data Analysis, 2nd Edition*. Chapman & Hall/CRC Text in Statistical Science.
- Gorczyński L (1920) Sur le calcul du degré du continentalisme et son application dans la climatologie. *Geografiska Annaler* 2: 324–331.
- Gorczyński L (1922) The calculation of the degree of continentality. *Monthly Weather Review* 7: 370.
- Goring S, Pellatt MG, Lacourse T, Walker IR and Mathewes RW (2009) A new methodology for reconstructing climate and vegetation from modern pollen assemblages: An example from British Columbia. *Journal of Biogeography* 36: 626–638.
- Guiot J, Torre F, Jolly D, Peyron O, Boreux J and Cheddadi R (2000) Inverse vegetation modelling by Monte Carlo sampling to reconstruct palaeoclimates under changed precipitation seasonality and CO<sub>2</sub> conditions: Application to glacial climate in the Mediterranean region. *Ecological Modelling* 127: 119–140.
- Haslett J, Whitley M, Bhattacharya S, Mitchell F, Allen J, Huntley B et al. (2006) Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society, Series A* 169: 395–438.
- Heinrichs ML, Walker IR and Mathewes RW (2001) Chironomid-based paleosalinity records in southern British Columbia, Canada: A comparison of transfer functions. *Journal of Paleolimnology* 26: 147–159.
- Hijmans RJ, Cameron SE, Parra JL, Jones PG and Jarvis AJ (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.
- Hinzman LD, Bettez ND, Bolton WR, et al. (2004) Evidence and implications of recent climate change in northern Alaska and other arctic regions. *Climatic Change* 72: 251–298.
- Holden P, Mackay A and Simpson G (2008) A Bayesian palaeoenvironmental transfer function model for acidified lakes. *Journal of Paleolimnology* 39: 551–566.
- Jackson ST, Betancourt JL, Booth RK and Gray ST (2009) Ecology and the ratchet of events: Climate variability, niche dimensions, and species distributions. *Proceedings of the National Academy of Sciences of the United States of America* 106: 19685–19692.
- Jones VJ, Solovieva N, Self AE, McGowan S, Rosén P, Salonen JS et al. (2011) The influence of Holocene tree-line advance and retreat on an arctic lake ecosystem; A multi-proxy study from Kharine Lake, North Eastern European Russia. *Journal of Paleolimnology* 46: 123–137.
- Juggins S (2007) *C2 Version 1.5 User Guide. Software for Ecological and Palaeoecological Data Analysis and Visualisation*. Newcastle upon Tyne: Newcastle University.
- Khotinskiy NA (1984) Holocene vegetation history. In: Velichko AA (ed.) *Late Quaternary Environments of the Soviet Union*. London: Longman, 179–200.
- Korhola A, Vasko K, Toivonen HTT and Olander H (2002) Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling. *Quaternary Science Reviews* 21: 1841–1860.
- Köster D, Racca JM and Pienitz R (2004) Diatom-based inference models and reconstructions revisited: Methods and transformations. *Journal of Paleolimnology* 32: 233–246.

- Kriticos DJ and Leriche A (2010) The effect of climate data precision on fitting and projecting species niche models. *Ecography* 33: 115–127.
- Kühl N, Moschen R, Wagner S, Brewer S and Peyron O (2010) A multiproxy record of late Holocene natural and anthropogenic environmental change from the *Sphagnum* peat bog Dürres Maar, Germany: Implications for quantitative climate reconstructions based on pollen. *Journal of Quaternary Science* 25: 675–688.
- Kultti S, Oksanen P and Välranta M (2004) Holocene tree line, permafrost, and climate dynamics in the Nenets Region, East European Arctic. *Canadian Journal of Earth Sciences* 41: 1141–1158.
- Kultti S, Välranta M, Sarmaja-Korjonen K, Solovieva N, Virtanen T, Kauppila T et al. (2003) Palaeoecological evidence of changes in vegetation and climate during the Holocene in the pre-Polar Urals, northeast European Russia. *Journal of Quaternary Science* 18: 503–520.
- Legendre P (1993) Spatial autocorrelation: Trouble or new paradigm? *Ecology* 74: 1659–1673.
- Lotter AF, Birks HJB, Eicher U, Hofmann W, Schwander J and Wick L (2000) Younger Dryas and Allerød summer temperatures at Gerzensee (Switzerland) inferred from fossil pollen and cladoceran assemblages. *Palaeogeography, Palaeoclimatology, Palaeoecology* 159: 349–361.
- MacDonald GM, Velichko AA, Kremenetski CV, Borisova OK, Goleva AA, Andreev AA et al. (2000) Holocene treeline history and climate change across northern Eurasia. *Quaternary Research* 53: 302–311.
- Moore PD, Webb JA and Collinson ME (1991) *Pollen Analysis*. Oxford: Blackwell.
- Oksanen PO, Kuhry P and Alekseeva RN (2001) Holocene development of the Rogovaya River peat plateau, European Russian Arctic. *The Holocene* 11: 25–40.
- Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC et al. (2001) Terrestrial ecoregions of the world: A new map of life on Earth. *Bioscience* 51: 933–938.
- Overpeck JT, Webb T III and Prentice IC (1985) Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs. *Quaternary Research* 23: 87–108.
- Paterson AM, Cumming BF, Dixit SS and Smol JP (2002) The importance of model choice on pH inferences from scaled chrysophyte assemblages in North America. *Journal of Paleolimnology* 27: 379–391.
- Peterson AT and Nakazawa Y (2008) Environmental data sets matter in ecological niche modelling: An example with *Solenopsis invicta* and *Solenopsis richteri*. *Global Ecology and Biogeography* 17: 135–144.
- Peyron O, Goring S, Dormoy I, Kotthoff U, Pross J, de Beaulieu J-L et al. (2011) Holocene seasonality changes in the central Mediterranean region reconstructed from the pollen sequences of Lake Accesa (Italy) and Tenaghi Philippon (Greece). *The Holocene* 21: 131–146.
- Prentice IC (1980) Multidimensional scaling as a research tool in Quaternary palynology: A review of theory and methods. *Review of Palaeobotany and Palynology* 31: 71–104.
- R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rekacewicz P (1998) *Ecosystems of Northwest Russia*. Barentswatch Atlas, GRID-Arendal, United Nations Environment Programme. Accessed from [http://maps.grida.no/go/graphic/ecosystems\\_in\\_northwest\\_russia](http://maps.grida.no/go/graphic/ecosystems_in_northwest_russia) on 3 April 2009.
- Robert CP and Casella G (2004) *Monte Carlo Statistical Methods*. New York: Springer-Verlag.
- Salonen JS, Seppä H, Välranta M, Jones VJ, Self A, Heikkilä M et al. (2011) The Holocene thermal maximum and late-Holocene cooling in the tundra of NE European Russia. *Quaternary Research* 75: 501–511.
- Sawada M, Viau AE, Vettoretti G, Peltier WR and Gajewski K (2004) Comparison of North-American pollen-based temperature and global lake-status with CCCma and AGCM2 output at 6 ka. *Quaternary Science Reviews* 23: 225–244.
- Seppä H and Bennett KD (2003) Quaternary pollen analysis: Recent progress in palaeoecology and palaeoclimatology. *Progress in Physical Geography* 27: 548–589.
- Seppä H and Birks HJB (2001) July mean temperature and annual precipitation trends during the Holocene in the Fennoscandian tree-line area: Pollen-based climate reconstructions. *The Holocene* 11: 527–537.
- Seppä H, Birks HJB, Odland A, Poska A and Veski S (2004) A modern pollen–climate calibration set from northern Europe: Developing and testing a tool for palaeoclimatological reconstructions. *Journal of Biogeography* 31: 251–267.
- Snow G (2008) TeachingDemos: Demonstration for teaching and learning. R package version 2.7. Accessed from <http://cran.rproject.org/web/packages/TeachingDemos/index.html> on 27 June 2011.
- St Jacques J-M, Cumming BF and Smol JP (2008) A 900-year pollen-inferred temperature and effective moisture record from varved Lake Mina, west-central Minnesota, USA. *Quaternary Science Reviews* 27: 781–796.
- Telford RJ and Birks HJB (2005) The secret assumption of transfer functions: Problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews* 24: 2173–2179.
- Telford R (2011) palaeoSig: Signifcons. Accessed from <http://cran.r-project.org/web/packages/palaeoSig/index.html> on 6 October 2011.
- Telford R (2011) palaeoSig: Significance tests for palaeoenvironmental reconstructions. Accessed from <http://cran.r-project.org/web/packages/palaeoSig/index.html> on 6 October 2011.
- Telford RJ and Birks HJB (2009) Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews* 28: 1309–1316.
- Telford RJ and Birks HJB (2011a) A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages. *Quaternary Science Reviews* 30: 1272–1278.
- Telford RJ and Birks HJB (2011b) Effect of uneven sampling along an environmental gradient on transfer-function performance. *Journal of Paleolimnology* 46: 99–106.
- ter Braak CJF and Juggins S (1993) Weighted averaging partial least squares regression (WA-PLS): An improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* 269/270: 485–502.
- Toivonen HT, Mannila H, Korhola A and Olander H (2001) Applying Bayesian statistics to organism-based environmental reconstruction. *Ecological Applications* 11: 618–630.
- Välranta M, Kaakinen A and Kuhry P (2003) Holocene climate and landscape evolution East of the Pechora Delta, East-European Russian Arctic. *Quaternary Research* 59: 335–344.
- Välranta M, Kaakinen A, Kuhry P, Kultti S, Salonen JS and Seppä H (2011) Scattered late-glacial and early-Holocene tree populations as dispersal nuclei for forest development in NE European Russia. *Journal of Biogeography* 38: 922–932.
- van der Voet H (1994) Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems* 25: 313–323.
- Vasko K, Toivonen HT and Korhola A (2000) A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction. *Journal of Paleolimnology* 24: 243–250.
- Velle G, Kongshavn K and Birks HJB (2011) Minimizing the edge-effect in environmental reconstructions by trimming the calibration set: Chironomid-inferred temperatures from Spitsbergen. *The Holocene* 21: 417–430.
- Xu Q, Li Y, Bunting MJ, Tian F and Liu J (2010) The effects of training set selection on the relationship between pollen assemblages and climate parameters: Implications for reconstructing past climate. *Palaeogeography, Palaeoclimatology, Palaeoecology* 289: 123–133.