Stomatal-based inference models for reconstruction of atmospheric CO₂ concentration: a method assessment using a calibration and validation approach

W. Finsinger* and F. Wagner-Cremer

(Palaeoecology, Institute of Environmental Biology, University of Utrecht, Laboratory of Palaeobotany and Palynology, Budapestlaan 4, NL-3584 CD Utrecht, The Netherlands)

Received 18 August 2008; revised manuscript accepted 9 February 2009

Abstract: We investigated changes in atmospheric CO₂ concentration (hereafter [CO₂]) over the period AD 1700–2002 as reconstructed using the inverse relationship between stomatal frequencies (SF) of Betula nana leaves from northern Europe and [CO₂]. The predictive ability of SF-inference models was assessed using a method of independent validation that involves two steps: (1) a training set of leaves grown between AD 1843 and 2002 was used to generate inference models; (2) the models were then applied to a fossil SF record of leaves grown between AD 1700 and 2002 that was split into two parts, a validation period (after AD 1850) and a reconstruction period (AD 1700–1850). Although our inference models had uncertainties comparable with other SF-inference models (root mean square error (RMSE) = c. 18–19 ppmv), uncertainties arising from the independent validation were larger (RMSE = c. 31–34 ppmv). Smoothed SF-inferred [CO₂] values after AD 1850 corresponded better to the industrial [CO₂] increase observed from instrumental records and from high-resolution ice cores, corroborating the accuracy of the reconstruction method in capturing a long-term (decadal- to centennial-scale) signal. This also indicates that in our record higher-frequency signals (eg, [CO₂] maxima around AD 1750) are potentially less reliable. In an attempt to estimate the maximum reconstruction uncertainty (± 67 ppmv), we considered (i) the RMSE of the validation (validation error) and (ii) the maximum difference between reconstructions obtained with different inference models during the validation period (method error). We suggest that reconstruction uncertainties may be reduced by reducing the uncertainty of our inference models, with a sub-fossil record characterized by lower variability in the SF time series over the validation period, and by smoothing the reconstruction. This study shows that independent validation is an important step to assess the precision and accuracy of quantitative proxy-based reconstructions.

Key words: Stomatal frequency, quantitative reconstructions, calibration, validation, climate change, Sweden.

Introduction

Stomata are small pores located on the surface of plant leaves and they play an important role in the regulation of the uptake of CO₂ for photosynthesis and the loss of water during transpiration. Although the roles of plant-growth regulators and environmental signals in regulating stomatal number are still largely unexplored (Bergmann and Sack, 2007), evidence shows that the concentration of atmospheric CO₂ (hereafter [CO₂]) controls not only stomatal opening (Young et al., 2006) but also stomatal frequencies (SF, eg, Woodward, 1987; Beerling et al., 1998). As shown by recent investigations, SFs are very likely regulated via systemic genetic-signalling pathways (Gray et al., 2000; Lake et al., 2002).

Because SFs are strongly influenced by [CO₂] they have the potential to be used for reconstructing past [CO₂] quantitatively (Roth-Nebelsick, 2005). The cuticle covering the leaf epidermis is made of waxes that are resistant to degradation under anoxic conditions, and thus it mirrors preserved morphological traits of the leaf epidermis inclusive stomata in sedimentary records. Transfer functions have been developed for a range of plant species to reconstruct [CO₂] based on SF (eg, Beerling et al.,...
Material and methods

Training-set leaves

The training set combines 23 *B. nana* and 14 *Betula pubescens* leaf samples, and covers a total of 34 years from the period 1843 to 2007.

* B. nana leaves collected in years 1919–1965 originate from specimens stored in the Nationaal Herbarium Nederland (NHN, Utrecht Universiteit branch) and leaves collected in years 1997–2002 originate from mature specimens growing at the Rasssejohka garden of the Forest Line Arboretum near the sub-Arctic Research Station Kevo (69°45′N, 27°E, Finland). In addition, *B. nana* leaves picked from a peat monolith that was cut from a mire surface near Kevo and sampled contiguously and at high-resolution (Hicks et al., 2004) were used to fill the gap between 1965 and 1997. The δ13C dating of individual *B. nana* leaves and *Sphagnum* leaves, stems and branches from this peat monolith have been wiggle-matched with the atmospheric δ13C content across the 1960s bomb peak (Goslar et al., 2005). Based on this, the rate of peat accumulation in the uppermost layers is in excess of 5 mm/yr, so each 2 mm slice is less than 1 year’s accumulation. *B. nana* collected from the Kevo peat monolith are dated to years AD 1976 to 1996.

* B. pubescens leaves are either herbarium specimens from the National Museum of Denmark (growing between AD 1843 and 1964), or modern leaves collected in northern Fennoscandia during AD 1997 to 2001. All training-set leaves originate from lowland sites (<200 m a.s.l.) in Fennoscandia between 54° and 69°N. For each year (n = 34), three to ten leaves were analysed.

For the modern training-set herbarium specimens of *B. nana* and *B. pubescens* were combined to test the comparability of the response rate and limit of these two species, which commonly hybridize to the mountain birch *B. pubescens ssp. czerepanovii* in northern Scandinavia (Wiegolski, 2005). The likelihood presence of leaf fragments of mountain birch in the subfossil assemblages can be counterbalanced by using the training set developed on the parent species of *B. pubescens ssp. czerepanovii* (Wagner et al., 2000).

Long stomatal density and stomatal index record

Determination of the age of samples in the Kiruna profile is based on radiocarbon-dated samples consisting of *Sphagnum* stems (Barneckow et al., 2007). Six of the ten samples analysed were deposited immediately prior to, or following, the atomic weapons testing of the 1960s, enabling us to assign calendar ages to the radiocarbon dates with high levels of precision. Based on these control points, an age-depth model was constructed (after Goslar et al., 2005) that was used to subsequently subsample the frozen monolith at an ‘annual’ resolution, assuming uniform peat accumulation between the control points (Figure 1a). The resulting sample-integration times vary from ~1 year (AD 2003–1955) to 3–7 years (AD 1954–1721) (Figure 1a).

Cuticle analysis

Subfossil leaf fragments of *B. nana* were separated from peat and identified by leaf and cuticle characteristics under a light microscope. For microscopic analysis, all leaf fragments were bleached in a cold 2–4% NaClO solution, washed in water and mounted on slides. Microscopic, computer-aided analysis of epidermal parameters was performed according to standard procedures on a Leica Quantimet 500C/500+ image-analysis system on seven digital images with a field size of 0.035 mm² (Wagner et al., 1996). On stomata-bearing aleuroles epidermal cell density (ED (n/mm²)) and stomatal density (SD (n/mm²)) were analysed and the area-independent stomatal index (SI(%)) = SD/(SD + ED) × 100) was calculated, following Salisbury (1927).
Atmospheric [CO₂] data
Historical [CO₂] data used for calibration are annual mean values as measured at Mauna Loa (AD 1958–2002) (Keeling and Whorf, 2005) and linearly interpolated [CO₂] measurements from air bubbles trapped in Antarctic ice cores (AD 1850–1957) (MacFarling Meure et al., 2006) that overlap very well with the instrumental Mauna Loa record.

Inference models and statistical analyses
Inference models were developed by formalizing the relation between stomatal frequencies and historical [CO₂] at the year of growth. We used two different methods to develop inference models. In order to account for the non-linear response of stomatal frequencies to changing [CO₂] (e.g., Kürschner et al., 1997), SDₜ, SIₜ and the historical [CO₂] values were log-transformed before fitting a linear response curve through the data sets (models SIlog and SDlog) (Figure 2A and B). In addition, since the log-transformation and regression tends to underestimate the extremes, thus reducing the amplitude of changes in the reconstruction, we scaled (mean and variance equalization) the SDₜ and SIₜ timeseries to the same mean and variance of the historical [CO₂] record used for calibration (models SIsca and SDsca) (Figure 2C and D).

The performance of the models is evaluated based on (i) their root-mean-squared error (RMSE), (ii) the RMSE of the validation and (iii) the Reduction of Error statistic (RE). The latter was used to estimate the strength of the linear relationship between validation and observation. The RE statistic, which ranges from −∞ to +1, compares the mean-square error (MSE) of the reconstruction to the MSE of a reconstruction that is constant in time with a value equivalent to the sample mean for the calibration data. If the

---

**Figure 1** Cuticle parameters of *Betula nana* leaf fragments in the Kiruna profile. (A) Age–depth relationship (whiskers indicate 2σ of calibrated age ranges) and estimated sediment accumulation rate (continuous line), (B) mean stomatal index (SIₜ%), (C) mean stomatal density (SDₜ)

---

**Figure 2** (A) and (B) Continuous line: [CO₂] inference models based on linear regression of log-transformed stomatal density (SDsca) (A) and stomatal index (SIₜ) data (B) for *Betula pubescens/nana*. (C) and (D): Continuous line: [CO₂] inference models based on scaling of stomatal density (SDsca) (C) and stomatal index (SIₜ) (D) data for *Betula pubescens/nana*. *Betula pubescens* (open circles) and *Betula nana* leaves (open triangles). Dashed lines indicate ± 1 RMSE.
reconstruction has any predictive value, one would expect it to do better than just the sample average over the calibration period; that is, one would expect RE to be greater than zero (see North et al., 2006).

### Results

#### Modern training set

The mean SD and SI of the training set (SD$_d$ and SI$_d$) over the [CO$_2$] range from 284 ppmv to 373 ppmv changed by almost 50% from 280 n/mm$^2$ (14 SI%) to 148 n/mm$^2$ (6.7 SI%). These changes largely exceed the mean standard deviation of sampled leaves from individual years, reflecting (i) the plastic response of the two species studied to changes in [CO$_2$] and (ii) the potential of the training set to serve for reconstructions of past [CO$_2$]. The smallest RMSE was found for the SD$_{log}$ model (16.71 ppmv CO$_2$) while other models (SI$_{log}$, SI$_{sca}$ and SD$_{sca}$) are characterized by RMSEs in the range of 18–19 ppmv CO$_2$ (Table 1).

#### Subfossil record

Over the period AD 1850–2003, subfossil $B.$ nana leaves show a successive long-term reduction in fossil-stomatal frequencies (SD$_F$ and SI$_F$) (Figure 2b and c). Mean SD$_F$ (SI$_F$) in the validation period (AD 1850–2002) range from 312 n/mm$^2$ (12.0%) to 86 n/mm$^2$ (5.6%). Before AD 1850, few data points of SI$_F$ are available because of the poor preservation of subfossil cuticles and only SD$_F$ were measured. SD$_F$ values before AD 1850 range from 174 n/mm$^2$ to 293 n/mm$^2$.

### Table 1 Performances of stomatal-based [CO$_2$] inference models for the calibration period (AD 1843–2002) and for the validation period (AD 1851–2000)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>$r^2$</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>log/log</td>
<td>0.68</td>
</tr>
<tr>
<td>scaling</td>
<td>−</td>
<td>18.02</td>
</tr>
<tr>
<td>SI</td>
<td>log/log</td>
<td>0.64</td>
</tr>
<tr>
<td>scaling</td>
<td>−</td>
<td>18.73</td>
</tr>
</tbody>
</table>

### Figure 3 Comparison between historical record of [CO$_2$] changes (Mauna Loa record and Law Dome record: continuous lines) and SF-inferred [CO$_2$] reconstructions obtained with different inference models. In order to highlight decadal-scale changes in SF-inferred [CO$_2$] changes, the record has been smoothed with a 3-point and with a 5-point moving average (dashed lines and dotted lines, respectively).
Validation of inference models
The performance of inference models was investigated, applying them to the SΔ and SI records. The SFσ-inferred reconstructions reflect the observed long-term trend of [CO₂] over the period AD 1850–2002 as recorded by high-resolution ice cores and by the instrumental record (Figure 3). However, the low reconstruction skill is indicated by higher RMSEs than in the calibration set and by negative RE values (Table 1). In an attempt to investigate the performance of the inference models in capturing longer-term trends of [CO₂] changes, n-point running means (n = 3 and n = 5) were calculated and predictions compared with the unsmoothed observed values. By doing this, RMSEs of smoothed reconstructions obtained with the SΔσ and SΔσσ models substantially decreased (Table 1). Notably, RMSEs of SD models were similar or even lower than those of inference models and REs considerably improved. In terms of RMSEs, reconstructions obtained with SIσ and SIσσ models did not improve much, while REs of smoothed SIσσ-inferred [CO₂] reconstructions remained low. Despite the improvements obtained by smoothing, SΔσσ-inferred [CO₂] are overestimated in parts of the records (eg, around AD 1900) or underestimated (eg, AD 1850–1880) (Figure 3).

Reconstruction of past [CO₂]
Because of the lack of SIσ values before AD 1850 and the better reconstruction skill of SΔσ and SΔσσ models, an attempt was made to reconstruct [CO₂] for the period AD 1850–1700 using the SΔσ record. SΔσ values in the reconstruction period range from 174 n/mm² to 293 n/mm², ie, within the range of SΔσ values used for the development of inference models (280 n/mm² to 148 n/mm²). Thus, reconstructions are not based on extrapolation of the models into the unknown, which may increase the reconstruction uncertainty (North et al., 2006).

In an attempt to estimate reconstruction uncertainties, two types of error were identified: uncertainty arising from (i) the different calibration methods that can be chosen for modelling the relationship between SF and historical [CO₂] (method error), and (ii) the unexplained variance in the reconstructions derived from validation against historical [CO₂] (validation error). The range between the maximum and minimum values for the reconstructions during the validation period is utilized to represent the method error. The validation error is estimated by means of the RMSE. These two error terms were quantified separately and combined (Figure 4) to allow for an estimation of the overall error of the historical [CO₂]. The combined uncertainty is represented by the square root of the summed and squared individual errors, following Esper et al. (2007). For the unsmoothed validation data set, the combined uncertainty is about ± 68 ppmv and decreases with increasing smoothing. For example, with a three-point smoothed data set the combined uncertainty reduces by c. 50% and equals c. ± 36 ppmv.

Reconstructed [CO₂] is characterized by high-amplitude variability (average amplitude is 70–100 ppmv). Lowest values are inferred with the SΔσσ model that does not overestimate low [CO₂] as the SΔσσσ does (Figures 3 and 4). Prominent minima in [CO₂] of about 260–270 ppmv occur around AD 1830, AD 1750 and AD 1720. In between these minima, [CO₂] maxima of up to 340–350 ppmv around AD 1780 and AD 1740 are reconstructed (Figure 4).

Discussion
[CO₂] controls both stomatal opening (Young et al., 2006) and stomatal numbers (SD and SI) in a wide variety of plants (Woodward, 1987; Beerling et al., 1998). However, SD also changes with light intensity (Salisbury, 1927; Tschá, 1982; Poole and Kürschner, 1999) and water stress (Tschá, 1982) through lateral cell expansion of the epidermal cells surrounding the stomata.
reconstructions of a variable of interest because all (paleoenvironmental) reconstruction procedures will produce a result in the form of computer output (Birks, 1998). Here we evaluate SdF-inferred [CO2] reconstructions with historical [CO2] records used to establish the inference models. Ideally, SdF-inferred [CO2] values should match with these historical [CO2] values. We observe that increasing SdF-inferred [CO2] after AD 1850 (Figure 3) corresponds well to the industrial [CO2] increase seen in instrumental and ice-core records (Keeling and Whorf, 2005). This clearly indicates that the higher amplitude of our SdF-inferred [CO2] record in comparison with ice-core records (Keeling and Whorf, 2005). On the other hand, lower RMSEs and higher REs of smoothed SFt-inferred [CO2] time series in the validation (Table 1) indicate that the apparent high-amplitude signals in the unsmoothed time series are less reliable. The higher variability of the unsmoothed SFt-inferred [CO2] time series may have three potential causes: first, the inherent variability in the phenotypic plasticity of the plants may contribute to a stochastic error in the reconstruction (e.g., RMSE of inference models). Second, the [CO2] record used to develop the inference models (Law Dome record) is itself partly a smoothed record (smoothing resulting from enclosure of air in the ice is about 10 years at DE08, Law Dome, Trudinger et al., 2003; MacFarling Meure et al., 2006). However, SdF-inferred [CO2] would suggest a higher variability of [CO2] changes than the annual instrumental record of Mauna Loa (Keeling and Whorf, 2005). This clearly indicates that the higher amplitude of our SdF-inferred [CO2] record in comparison with ice-core records cannot be related to the smoothing effect of processes involved in air trapping but also to uncertainties in the SFt-based reconstructions, as was previously suggested (e.g., McElwain et al., 2002; Van Hoof et al., 2005). Third, the variability of the SdF time-series is higher than that of the SdF series. We may hypothesize that the variability in the training set is lower because within-leaf variation of SF was minimized by means of a well-defined sampling strategy (Poole et al., 1996). Subfossil leaves from the Kiruna profile were rarely complete and fragments analysed could not always be assigned to the central leaf area. This potentially influences the estimate of the combined reconstruction uncertainty, which appears to be c. three times as large as the RMSE of the inference models.

**SdF-inferred [CO2] changes AD 1700–1850**

Knowledge of past [CO2] variations is of utmost importance for our understanding of global biogeochemical cycles and for climate modelling. The most direct way of investigating past variations of [CO2] before AD 1958, when continuous [CO2] measurements started (Keeling and Whorf, 2005), is the analysis of air extracted from suitable ice cores. Detailed [CO2] records from Antarctica (e.g., Dronning Maud (DML, Siegenthaler et al., 2005) and Law Dome (MacFarling Meure et al., 2006)) indicate that between AD 1000 and 1850 variations of up to 2 ppmv occurred. Therefore it is generally assumed that climate changes prior to AD 1250 are uncertain and may have been slightly larger than represented in ice core–air records (Trudinger et al., 2003; MacFarling Meure et al., 2006). In fact, air in a sample of ice does not correspond to a single time in the past, but is a mix of air parcels that left the atmosphere over a range of times (Trudinger et al., 2003). The main processes involved in the air trapping are (1) diffusion through the firm layer and (2) the gradual trapping of air into bubbles in ice in the lock-in zone (Trudinger et al., 2003; Siegenthaler et al., 2005). The age spread from bubble close off is largely determined by the snow-accumulation rate, with higher accumulation-rate sites having narrower age distributions than low accumulation-rate sites. For three high-resolution ice-core records spanning the time interval AD 1700–present, the accumulation rate is highest at DE08 (Law Dome record AD 1844–1973, Etheridge et al., 1996; MacFarling Meure et al., 2006). At DE08 the smoothing corresponds broadly to a 10-year running mean (Trudinger et al., 2003) and at the DSS site (Law Dome record: AD 1–1884, MacFarling Meure et al., 2006) the age distribution for [CO2] in ice is almost certainly wider than at DE08 (Trudinger et al., 2003). In comparison with the DSS site, the accumulation rate at EDML is assumed very low (Siegenthaler et al., 2005). Hence, if the amplitude and the rapidity of SdF-inferred [CO2] changes between AD 1750 and 1850 were real, they should better match with [CO2] changes recorded at Law Dome (DSS site).

It is notable that between AD 1700 and 1850, mean SdF-inferred [CO2] are generally higher than mean [CO2] values estimated from ice-core records, which range between 276 and 288 ppmv (Figure 4). Consistently higher SF-inferred [CO2] than ice-core [CO2] records have been observed for the Holocene by McElwain et al. (2002), who related the discrepancy to (i) problems with the transfer functions used to predict [CO2] from past SF changes or to (ii) the influence of environmental variables on SD. We can exclude that high [CO2] in the reconstruction are due to the transfer function alone because lowest values were correctly predicted with the use of the SFt model in the validation (possibly underestimated with the Sd model). The influence of other environmental factors (e.g., changing light intensity and water stress) are minimized (see discussion above).

In addition to generally higher values in comparison with ice-core [CO2] records, our SdF-inferred reconstruction also shows two maxima of larger magnitude (c. 60–70 ppmv) and of short duration (less than 100 years) centred between AD 1700 and 1800 (Figure 4). A prominent maximum (within uncertainty bands) similar in timing (considering age uncertainties, Figure 1) and in magnitude centred at around AD 1750 was recorded in an Sd-inferred [CO2] reconstruction based on *T. heterophylla* needles from northeastern America by Kowuenberg et al. (2005) (Figure 5). Because the [CO2] variation also shows similarities with terrestrial air temperature trends, Kowuenberg et al. (2005) suggested that [CO2] could have served as a forcing factor for climate change also prior to the mid-nineteenth century. By contrast, a *Salix herbacea* SI-inferred [CO2] record from Fennoscandia (not shown) indicated no significant change in reconstructed values between AD 1700 and 1850 (Rundgren and Beerling, 1999). In the latter record, reconstructed values were within the 95% confidence limits of the reconstruction (c. ± 20 ppmv). In our new SdF-inferred [CO2] record, reconstruction uncertainties are larger than that (combined reconstruction uncertainty is c. ± 67 ppmv), and can therefore not confirm the conclusion of significantly higher [CO2] values during the period AD 1700–1850 with any level of confidence (Figure 4). The reconstruction uncertainty is almost as large as the [CO2] changes that occurred during glacial terminations as revealed by analysis of CO2 trapped in Antarctic ice cores (Lüthi et al., 2008). It is essential to reduce the reconstruction uncertainty. This may be achieved in at least two ways: (i) the reduction of the uncertainty of our inference models by adding more samples used in the calibration of the proxy (e.g., Lucy et al., 2008), and (ii) with a subfossil record characterized by lower variability in the SF time series over the validation period. The latter may be achieved if a record with higher abundance of well-preserved cuticles and complete leaves can be found. In addition, a higher datapoint density in the reconstruction period would enable the smoothing of the record thereby reducing the uncertainties of the high-frequency record.
Conclusions

With ever-increasing interest in past climate changes, it is important to assess the confidence in the quality of proxy-based reconstructions of environmental variables. Here we attempt to investigate the performance of inference models for the prediction of [CO₂] changes based on subfossil stomata of Betula nana leaves. We generated a training set with leaves collected from herbariums and from a short peat section and historical [CO₂] records from instrumental and ice-core records. Subsequently, inference models were applied to an independent subfossil SF record. The latter was obtained from a sediment core deposited over the time span from AD 1700 to 2003 that was accurately sampled allowing for a near-annual resolution between AD 1950 and present and a c. 5 year resolution for older samples. This well-dated record enabled the independent validation of SF-based [CO₂] reconstructions by comparison between observed versus predicted [CO₂] values showing that (i) scaling-based inference models perform better than regression-based models, (ii) the reconstruction uncertainty is larger (by a factor of two) than the uncertainty arising from the training set, and (iii) that the long-term (decadal- to centennial-scale) changes are better retained than the higher frequency changes. Hence, although our SDₐ-inferred [CO₂] reconstruction resembles in its long-term temporal pattern another SF-inferred [CO₂] record from northeastern America (Kouwenberg et al., 2005), it is clearly essential to reduce the reconstruction uncertainty to enhance the accuracy and precision of stomatal-based [CO₂] reconstructions.

Acknowledgements

We are grateful to Timme Donders, Hiro Yamazaki and Danny McCarroll for discussions and to two anonymous reviewers for constructive comments. Atmospheric [CO₂] from Law Dome and South Pole ice cores were downloaded from the PANGAEA database (http://dx.doi.org/10.1594/PANGAEA.472477 and http://dx.doi.org/10.1594/PANGAEA.472477). WF was financially supported by the European Union 017008 (GOC) ‘MILLENIUM’ project. This is Netherlands Research School of Sedimentary Geology (NSG) publication no. 20090403.

Figure 5 Normalized SF-inferred [CO₂] records (zero mean and unit standard deviation) from Jay Bath (Kouwenberg et al., 2005) and from the Kiruna record (see Figure 4)

References


Garcia-Amorena, I., Wagner, F., van Hoof, T.B. and Gomez-Manzaneque, F. 2006: Stomatal responses in deciduous oaks from southern Europe to the anthropogenic atmospheric CO₂ increase; refining the stomatal-based CO₂ proxy. Review of Palaeobotany and Palynology 141, 303–12.


Hicks, S., Goslar, T. and van der Borg, K. 2004: A near annual record of recent tree line dynamics from Northern Finland. *Acta Palaeobotanica* 223–224, 5–11.


