

# A matter of dispersal: REVEALSinR introduces state-of-the-art dispersal models to quantitative vegetation reconstruction

Martin Theuerkauf<sup>1,3</sup> · John Couwenberg<sup>2,3</sup> · Anna Kuparinen<sup>4</sup> · Volkmar Liebscher<sup>5</sup>

Received: 30 December 2015 / Accepted: 1 April 2016 / Published online: 13 April 2016  
© Springer-Verlag Berlin Heidelberg 2016

**Abstract** The REVEALS model is applied in quantitative vegetation reconstruction to translate pollen percentage data from large lakes and peatlands into regional vegetation composition. The model was first presented in 2007 and has gained increasing attention. It is a core element of the Landcover 6k initiative within the PAGES project. The REVEALS model has two critical components: the pollen dispersal model and pollen productivity estimates (PPEs). To study the consequences of model settings, we implemented REVEALS in R. We use a state-of-the-art Lagrangian stochastic dispersal model (LSM) and compare model outcomes with calculations based on a conventional Gaussian plume dispersal model (GPM). In the LSM turbulence causes pollen fall speed to have little effect on the

dispersal pattern whereas fall speed is a major factor in the GPM. Dispersal models are also used to derive PPEs. The unrealistic GPM produces PPEs that do not describe actual pollen productivity, but rather serve as a basin specific correction factor. A test with pollen and vegetation data from NE Germany shows that REVEALS performs best when applied with the LSM. REVEALS applications with the GPM can produce realistic results, but only if unrealistic PPEs are used. We discuss the derivation of PPEs and further REVEALS applications. Our REVEALS implementation is freely available as the ‘REVEALSinR’ function within the R package DISCOVER. REVEALSinR offers an environment for experimentation and analysing model sensitivities. We encourage further experiments and welcome comments on our tool.

Communicated by F. Bittmann.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00334-016-0572-0) contains supplementary material, which is available to authorized users.

✉ Martin Theuerkauf  
martin.theuerkauf@uni-greifswald.de

- <sup>1</sup> Institute for Geography and Geology, Ernst-Moritz-Arndt-University Greifswald, Friedrich-Ludwig-Jahn-Straße 16A, 17487 Greifswald, Germany
- <sup>2</sup> Institute of Botany and Landscape Ecology, Ernst-Moritz-Arndt-University Greifswald, Soldmannstraße 15, 17487 Greifswald, Germany
- <sup>3</sup> Greifswald Mire Center, Soldmannstraße 15, 17487 Greifswald, Germany
- <sup>4</sup> Department of Environmental Sciences, University of Helsinki, P.O. Box 65, 00014 Helsinki, Finland
- <sup>5</sup> Institute for Mathematics and Informatics, Ernst-Moritz-Arndt-University of Greifswald, Walther-Rathenau-Straße 47, 17487 Greifswald, Germany

**Keywords** DISCOVER · Lagrangian stochastic models · Pollen · Fall speed of pollen · Pollen productivity estimates

## Introduction

Reconstructing past plant abundances from the pollen record is one of the main goals in palynology since the inception of the field some 100 years ago (von Post 1918). This goal is notoriously hard to achieve. The relationship between plant abundances and pollen is not straightforward, because different plant taxa produce different amounts of pollen that are dispersed with different effectiveness. Differences in dispersion interact with the production bias and lead to over- or underrepresentation of taxa in the pollen record: a taxon with very large, poorly dispersed pollen grains and low pollen productivity is obviously under-represented in the pollen record of a large lake. Yet, because its pollen travels shorter distances, the

same taxon may be over-represented in the pollen record of a small forest hollow.

Ad hoc attempts to correct over- and underrepresentation of plant taxa in the pollen record have a long history. The first well-known formalized approach is the R-value approach by Davis (1963), later refined to the extended R-value approach by Parsons and Prentice (1981). The R-value approach uses a taxon-specific correction factor (the ratio of R-values) to correct for production and dispersal bias *at the same time*. However, the example above illustrates that the representation of a taxon in the pollen record may differ between different basins. R-values are therefore not universal: they need to be calibrated separately for each basin type.

The REVEALS approach (Sugita 2007a) overcomes this limitation by correcting the production bias and the dispersal bias *separately*. It uses pollen productivity estimates (PPEs) to account for the production bias and pollen fall speeds and the associated ‘pollen dispersal-deposition coefficient’ or K-factor to account for the dispersal bias. PPEs ideally represent how much pollen a taxon produces in relation to a reference taxon. PPEs are estimated in studies that relate surface pollen deposition to distance weighted plant abundances in the surroundings of the pollen sample sites. Because distance weighting is achieved through application of a pollen dispersal model, the quality of PPEs depends on the suitability of the underlying dispersal model (Theuerkauf et al. 2013). Also the K-factor is calculated with a specific pollen dispersal model. It represents how much pollen of a taxon is deposited in a lake or peatland with a known diameter compared to the amount of pollen deposited in a basin with a zero diameter. The K-factor is 1 in a basin with zero diameter and declines with increasing basin size.

REVEALS has gained increasing attention over the past years; it is a core element of the Landcover 6k initiative within the PAGES project (<http://www.pages-igbp.org/ini/wg/landcover6k/>). REVEALS is also integral part of the landscape reconstruction algorithm (LRA), which aims at reconstructing vegetation composition on a local scale (Sugita 2007b). So far, all REVEALS applications rely on a Gaussian plume model (GPM) for pollen dispersion (Sutton 1947) in both the calibration of PPEs as well as in the REVEALS application itself. Recent developments, which will be outlined in the next section, have highlighted the limitations of this dispersal model family. Lagrangian stochastic models (LSM) describe pollen dispersion more realistically, especially when it comes to long-distance dispersal (Kuparinen et al. 2007). The better performance of LSMs has been demonstrated using surface pollen and modern vegetation data (Theuerkauf et al. 2013). We show how this progress in the modelling of pollen dispersal affects REVEALS reconstructions. For this purpose we

developed an implementation of the REVEALS model in the R environment for statistical computing (R Core Team 2013). As part of the DISCOVER package, ‘REVEALSinR’ is available as open source software.

### Dispersal models in palynology

Pollen dispersal models play a critical role in quantitative reconstructions of past vegetation. Reliable reconstructions of past vegetation require an understanding of where the pollen comes from. Despite its central role in vegetation reconstruction, the study of atmospheric dispersion of small particles such as pollen is covered by other fields of research, such as aerobiology, micrometeorology, the military (to study dispersion of radioactive substances or chemical weapons), medicine (to forecast hay fever potential), agriculture (to control pests or transgenic plants) and forestry (to assess pollination potentials). Pollen dispersal models developed during the 20th century can be categorized as follows:

- (i) Simple mathematical models with only few parameters that describe observed dispersal patterns in a correlative way (e.g. Schmidt 1918; Gregory 1945; Tauber 1965).
- (ii) Quasi-mechanistic models with descriptive parameters that are estimated by statistical fitting to empirical data (Tufto et al. 1997; Nurminiemi and Tufto 1998; Klein et al. 2003).
- (iii) Fully mechanistic models that describe the physical factors affecting dispersal and are therefore able to predict the dispersal process based on measurements of environmental parameters (Kuparinen 2006; Kuparinen et al. 2007; Theuerkauf et al. 2013).

The first to adopt dispersal models in pollen-based vegetation reconstruction was Tauber (1965). Later, also Prentice (1985) used the same equations of Sutton for a GPM (Sutton 1947, 1953) to calculate the origin of pollen in peatlands of different size. Sugita (2007a, b) incorporated this dispersal model in his landscape reconstruction algorithm (LRA). This model framework is designed to quantify regional and local scale past plant abundances using pollen data from large and small sites (see e.g. Hultberg et al. 2015; Mehl and Hjelle 2016). The LRA optionally adjusts the GPM of Sutton to pollen deposition in lakes.

Overall, simple dispersal models such as the GPM fail to predict the magnitude of long-distance dispersal (Kuparinen 2006). Field observations have indicated, for example, that cross-pollination and seed dispersal by wind commonly occur over much larger distances than predicted (Giddings et al. 1997; Hofmann et al. 2014). Experiments

and micrometeorological modelling both suggest that strong upward air sweeps, so-called ‘updrafts’ are a key driver of long-distance dispersal (Nathan et al. 2002; Tackenberg 2003). Updrafts lift airborne particles above the canopy where the horizontal airflows are stronger, dispersing particles over large spatial distances (Soons et al. 2004). Such turbulent events are generally not described by GPMs; if GPMs include turbulent flows then these are assumed to be symmetric, non-autocorrelated fluctuations around the mean horizontal airflow. Therefore, GPMs appear only suitable to predict dispersal at short distances (<15 m), because only over such short distances dispersal is governed by release height and mean wind speed rather than the turbulence conditions (Soons et al. 2004). Yet, even in closed forest hollows most pollen arrives from longer distances. The discrepancies between model outcome and observations have stimulated the development of new modelling approaches since the early 21st century.

Realistic models of long-distance dispersal of pollen and seeds have come to depend on Lagrangian stochastic simulations as the state-of-the-art tool (Kuparinen 2006; Nathan et al. 2011). LSMs predict the trajectory of each dispersing particle under turbulent conditions, which depend on the degree of atmospheric (in)stability and the vertical structure of the atmospheric boundary layer. Within the canopy, turbulence is weak and close to symmetric, while above it turbulence is characterized by strong upward sweeps and weaker, but more frequent downward flows (Kuparinen et al. 2007).

Intuitively, one might assume that atmospheric conditions have larger impact on pollen with low fall speed than on pollen with high fall speed. However, sensitivity analyses reveal the opposite: dispersal of pollen with low fall speed hardly depends on atmospheric conditions as its falling velocity is typically lower than average vertical turbulent flows. In contrast, dispersal of pollen with high fall speed depends on strong turbulent flows that are capable of carrying also such pollen across longer distances (Kuparinen et al. 2007). Pollen is primarily released under unstable atmospheric conditions with strong turbulent flows (Jackson and Lyford 1999) so that the difference in the dispersal of pollen is largely independent of fall speed. Strong updrafts under unstable conditions lift pollen both with low and high fall speeds well above the canopy, initiating long-distance transport (Soons et al. 2004).

Upland pollen deposited in large lakes or peatlands to a large degree arrives from some to many kilometres distance. Observing pollen dispersal over such distances to test dispersal models directly is virtually impossible. Dispersal models instead may be tested using modern pollen and vegetation data. A first such test on lakes across NE Germany has indeed shown that the LSM of Kuparinen

et al. (2007) much better describes observed pollen deposition than the GPM (Theuerkauf et al. 2013).

The GPM and LSM differ considerably in the predicted deposition from various sources (Fig. 1). The contribution of pollen arriving from 10 to 100 km away is much lower in the GPM than in the LSM. The LSM predicts that some 20–30 % of the pollen arriving from outside a peatland with a diameter of 1,000 m originates from within 10 km distance, for both lighter and heavier pollen types. Deposition of pollen from increasingly farther away gradually declines. Very little pollen is predicted to arrive from beyond 100 km. In contrast, the GPM (adjusted for neutral conditions) predicts that pollen arriving from the first 10 km is far more important; for heavier pollen making up close to 80 % of the total deposition. Consequently, the amount of pollen that arrives from greater distances is very low. Yet, the long tail of the Gaussian distribution means that a considerable amount of the deposited pollen comes from distances beyond 100 km, from up to thousands of kilometres away. In the GPM adjusted for unstable conditions deposition from nearby sources is somewhat lower, but deposition from very long distance is even higher.

Differences between the deposition of pollen with high and low fall speed are—as mentioned—small for the LSM but high for the GPM. For the centre of a peatland of 1,000 m diameter the LSM predicts that 80 % of total upland pollen deposition originates from within 60 km for pollen with low fall speed and 50 km for pollen with high fall speed (Table 1). The GPM for neutral conditions predicts that the size of the 80 % source area is 119.5 km for taxa with low fall speed and 12.2 km for pollen with high fall speed. The GPM for unstable conditions predicts far larger source areas.

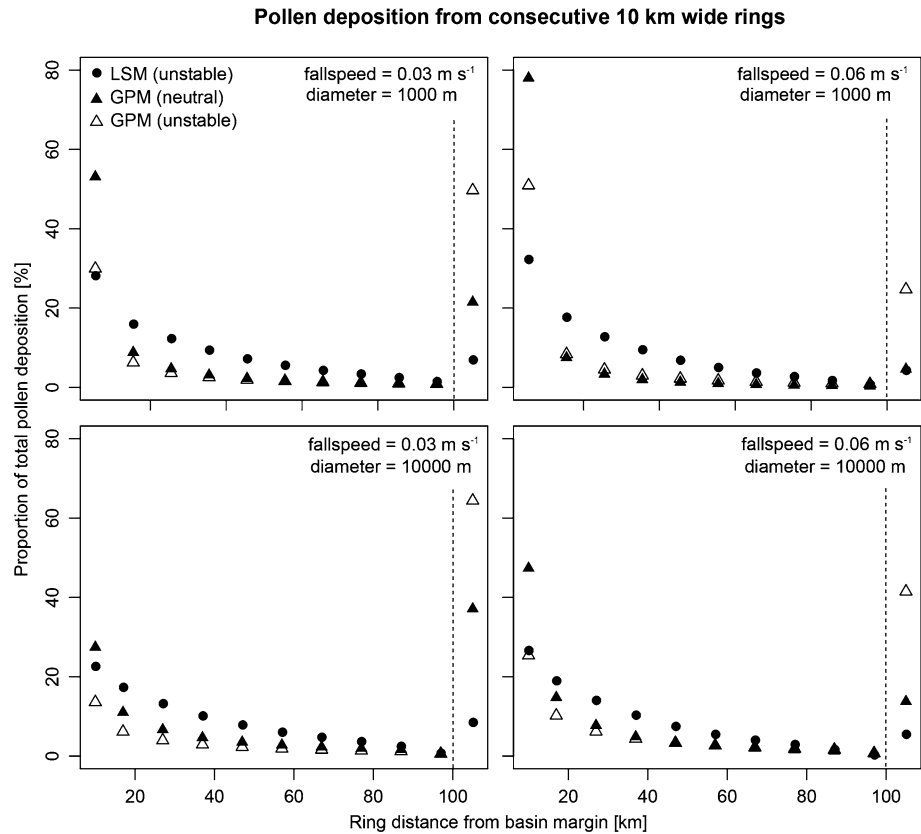
### Principles of ‘REVEALSinR’

The REVEALS model (Sugita 2007a) is based on the assumption that pollen deposition of a plant taxon in a large lake or peatland is equal to the mean abundance of that taxon in the region, multiplied by its pollen productivity and its ‘pollen dispersal-deposition coefficient’  $K$ . In reverse, if pollen data are available, the past regional abundance of a taxon can be calculated as its pollen deposition divided by its pollen productivity and dispersal coefficient. The REVEALS model expresses abundance in relative terms because pollen data are commonly given as percentage data:

$$V_i = 100 \times \frac{n_i/PPE_i K_i}{\sum_{j=1}^m n_j/PPE_j K_j}$$

where  $V_i$  is the relative abundance of taxon  $i$ ,  $n_i$  is the pollen count of  $i$ ,  $PPE_i$  is the pollen productivity estimate

**Fig. 1** Origin of upland pollen in a peatland with 1,000 m (above) and 10,000 m (below) diameter, calculated for low (left) and high fall speed of pollen (right). Origin of pollen is calculated for consecutive rings of 10 km width with the LSM and the GPM adjusted to neutral and unstable atmospheric conditions



**Table 1** Radius of the 80 % source area of pollen, i.e. the distance from which 80 % of the total pollen deposition at a site arrives. Radius calculated for taxa with low and high fall speed of pollen and

for deposition in peatland sites of different diameter using different dispersal models

Dispersal model	LSM		GPM neutral		GPM unstable	
	1,000	10,000	1,000	10,000	1,000	10,000
Basin diameter (m)						
Fall speed ( $\text{m s}^{-1}$ )	0.03	0.06	0.03	0.06	0.03	0.06
Radius (km) of the 80 % source area	60	50.5	67.3	56.8	119.5	12.2
					415	61.5
					>1,000	185
					>1,000	700

for  $i$ ,  $K_i$  is the ‘pollen dispersal-deposition coefficient’ for  $i$  and  $m$  is the total number of pollen types considered.

The REVEALS model was originally implemented in the C++ programming language by Shinya Sugita, (current version ‘v4.2.2.Tallinn.wks.exe’, Mazier et al. 2012). REVEALS calculates K-factors using a GPM adjusted to neutral atmospheric conditions, although adjustment to unstable conditions would be more appropriate (Jackson and Lyford 1999). The model offers an option for pollen deposition in lakes, taking account of lake internal mixing (Sugita 1993).

Our alternative implementation ‘REVEALSinR’ is written in the R environment for statistical computing (R Core Team 2013; see ESM for details). Conceptually, ‘REVEALSinR’ differs from the Sugita programme in the calculation of K-factors and in the calculation of error estimates. ‘REVEALSinR’ is flexible with respect to the

dispersal model used. By default, it uses a LSM, but GPMs (adjusted to unstable or neutral conditions) and the non-parametric function ‘1 over d’ are implemented as well. Alternative models can easily be added. Because actual LSM calculations are time consuming, ‘REVEALSinR’ uses look-up tables of LSM outputs that cover a range of fall speeds and atmospheric conditions.

Like the original REVEALS programme, ‘REVEALSinR’ includes a function to address deposition in lakes (for details see ESM). Both the original REVEALS programme and ‘REVEALSinR’ only consider atmospheric pollen deposition (and lake mixing); neither model is applicable to sites that receive significant amounts of pollen from rivers, streams or surface run-off.

In the original REVEALS programme error estimates are calculated from the variance–covariance matrix of PPEs

through a hybrid method (Sugita 2007a). ‘REVEALSinR’ arrives at error estimates through repeated model runs (a minimum of 1,000) with random error added in pollen data and PPEs during each model run (see ESM). By default, the 10 and 90 % percentile of the repeated calculations are selected as error range boundary estimates. Other options are easily implemented.

The ‘REVEALSinR’ function is freely available on our website at <http://discover.botanik.uni-greifswald.de>. The ‘REVEALSinR’ function is the first function of the DISCOVER package.

## Materials and methods

To introduce and test the ‘REVEALSinR’ function we first use a simple scenario with two taxa X and Y. Both taxa produce similar amounts of pollen (PPE = 5; SE = 0.5) but with different fall speeds: X has a higher fall speed ( $0.06 \text{ m s}^{-1}$ ) than Y ( $0.03 \text{ m s}^{-1}$ ). X and Y are similarly abundant in the pollen record: 500 pollen grains of each taxon are found. We associate the record with lakes and peatlands of different size (100–10,000 m in diameter), using different cut-off distances for the tail of the GPM (50 km to infinity). This cut-off sets an arbitrary limit to the maximum distance pollen may travel (the region considered as pollen source area). The cut-off for the LSM is set to 100 km, which is the calculated average distance at which 95 % of the pollen has settled (cf. Fig. 1). We calculate regional vegetation composition with ‘REVEALSinR’ using the LSM as well as the GPM. The LSM parameters apply to unstable atmospheric conditions (friction velocity  $u^* = 0.6 \text{ m s}^{-1}$ , Obukhov-length  $L = -40 \text{ m}$ ; further parameters follow Kuparinen et al. 2007). Sugita’s REVEALS programme uses the GPM with parameters for neutral atmospheric conditions (vertical diffusion coefficient  $c_z = 0.12$ , turbulence parameter  $n = 0.25$ ). We include this setting for comparison, but also use the GPM with parameters for unstable conditions ( $c_z = 0.21$ ,  $n = 0.20$ ).

Secondly, ‘REVEALSinR’ is applied to a high resolution pollen dataset from Lake Tiefer See/NE Germany covering the period 1870–2010 (Theuerkauf et al. 2015). Calculations are performed for four different settings A, B, C and D that differ in the underlying pollen dispersal model and PPE dataset (Table 2). A and B use the LSM, C and D the GPM; A and C use the PPE.MV2015 dataset, B and D the PPE.st2 dataset (Table 2). The PPE.MV2015 (Table 3) dataset was specifically derived for the study area of NE-Germany using the LSM (Theuerkauf et al. 2013, 2015). The PPE.st2 dataset (Table 3) has been compiled from a number of PPE studies across northern and central Europe (all using the GPM; Mazier et al. 2012). Further options, i.e. basin size (300 m), basin type (lake) and cut-off size of

the pollen source area (100 km), are equal in all experiments. Experiment C and D were repeated as C\* and D\* with Sugita’s REVEALS programme (latest version: REVEALS.v4.2.2.Tallinn.wks.exe). To validate model performance we compare the reconstructed cover of major crops during the study period with observed cover values recorded in written sources (cf. Theuerkauf et al. 2015). Cover data for trees are only available for the present so that model results are validated for the modern situation only. In the text, elements of the actual vegetation are written in italics, whereas reconstructed taxa are written in normal font.

## Results

### The two taxa scenarios

The vegetation composition that was calculated from the hypothetical pollen sample differs strongly depending on which dispersal model is used (Fig. 2). With the LSM, the 50 % of pollen of X (with a high fall speed) translates into a cover in the regional vegetation slightly above 50 %. Consequently, the cover of Y (with low fall speed) is slightly below 50 %. Whether the sample is assumed to be taken from a peatland or lake has little effect. Also the influence of basin size is small. The highest cover of X (52.5 %) is found for a peatland with a diameter of 1,000 m. With the GPM, the cover of X is instead modelled to be well above 70 % (and that of Y well below 30 %). The cover of X increases with basin size from 74.4 % for a lake and 75 % for a peatland with a diameter of 100 m towards 82.9 % for a lake and 85 % for a peatland with a diameter of 10,000 m; the cover of Y decreases correspondingly. The difference in cover between X and Y is larger for peatlands than for lakes.

The different dispersal functions (LSM and GPM) result in differences in the K-factors (=relative pollen influx) of the models. The K-factor is 1 for basins of zero size and decreases with increasing basin size as expected (Fig. 2). However, the decrease is much stronger with the GPM than with the LSM. In other words, K-factors for the LSM are much higher (0.5–0.8) than for the GPM (0–0.3), meaning that for medium sized to large basins the LSM predicts significantly higher pollen deposition arriving from within the 100 km region than the GPM. Moreover, with the LSM K-factors for X and Y hardly differ, whereas with the GPM the K-factors for taxon Y with light pollen are 2–5 times higher than for taxon X with heavier pollen. The ratio of Y:X increases with basin diameter and is higher for peatlands than for lakes.

Increasing the size of the region considered as pollen source area (i.e. cutting the tail of the GPM at a larger distance) increases the K-factor of Y (with the lower fall

**Table 2** Studied model settings; radius of the source area considered is 100 km, the basin type is lake

Setting	A	B	C	D	C*	D*
Software	REVEALSinR	REVEALSinR	REVEALSinR	REVEALSinR	Sugita progr.	Sugita progr.
PPE	PPE.MV2015	PPE.st2	PPE.MV2015	PPE.st2	PPE.MV2015	PPE.st2
Dispersal model	LSM	LSM	GPM	GPM	GPM	GPM

**Table 3** Fall speed of pollen, pollen productivity estimates and their error from the PPE.st2 dataset (calculated with the GPM) and the PPE.MV2015 dataset (calculated with the LSM). The PPE.MV2015 dataset does not include PPEs for *Acer*, *Carpinus*, *Corylus*, *Salix*, *Calluna*, Cerealia and Cyperaceae; these values were partly taken from the PPE.st2 dataset. The PPE of *Corylus* was set to 10, assuming that pollen productivity is about as high as in *Betula*. For Cerealia, the mean PPE from the period 1950–2010 from the lake Tiefer See data is used. The Cerealia PPE includes *Secale* in PPE.st2 but excludes *Secale* in PPE.MV2015

Taxon	Fall speed	PPE.st2	PPE.st2 SE	PPE.MV2015	PPE.MV2015 SE
<i>Acer</i>	0.056	1.27	0.23	1.27	0.23
<i>Alnus</i>	0.021	9.07	0.10	16.26	1.24
<i>Betula</i>	0.024	3.09	0.27	13.77	2.96
<i>Carpinus</i>	0.042	3.55	0.43	3.55	0.43
<i>Corylus</i>	0.025	1.99	0.20	10.00	1.00
<i>Fagus</i>	0.057	2.35	0.11	2.20	0.21
<i>Fraxinus</i>	0.022	1.03	0.11	1.07	0.31
<i>Picea</i>	0.056	2.62	0.12	0.87	0.18
<i>Pinus</i>	0.031	6.38	0.45	4.93	0.22
<i>Quercus</i>	0.035	5.83	0.15	9.15	1.24
<i>Tilia</i>	0.032	0.80	0.03	2.57	2.56
<i>Ulmus</i>	0.032	1.27	0.05	2.60	2.55
<i>Salix</i>	0.022	1.22	0.11	1.22	0.11
<i>Artemisia</i>	0.025	3.48	0.20	4.56	0.51
<i>Calluna</i>	0.038	0.82	0.02	0.82	0.02
Cerealia	0.06	1.85	0.38	0.15	0.02
Cyperaceae	0.035	0.87	0.06	0.87	0.06
Poaceae	0.035	1.00	0.00	1.00	0.10
<i>Plantago lanceolata</i>	0.029	1.04	0.09	2.98	1.08
<i>Rumex acetosella</i>	0.018	2.14	0.28	3.55	0.56
<i>Secale</i>	0.06	3.02	0.05	1.17	0.14

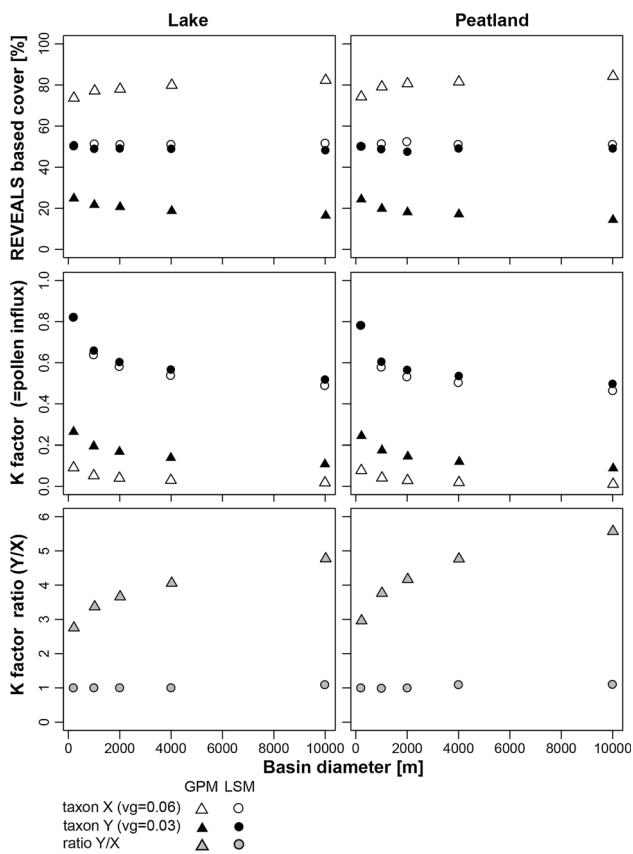
speed) far more than that of X (Fig. 3). As a result, the reconstructed cover of Y decreases. The effect is similar in basins with different diameter but stronger with the GPM adjusted for unstable conditions than with the GPM adjusted for neutral conditions.

### Lake Tiefer See

Analysis of the Tiefer See data showed that among the six model settings (Table 2), setting A produces the best fit between the REVEALS based plant cover reconstructions and observed plant cover (Table 4). With this setting A, which uses the LSM and PPE.MV2015, the reconstructed cover of cereals (excluding *Secale*), grassland and *Secale* largely matches the observed cover over the study period (Fig. 4). Deviations occur particularly for the 1970s and onward, for Cerealia also before. The setting also produces a good fit for *Alnus* but a somewhat too high cover for *Fagus*,

*Picea* and *Pinus* (Fig. 5). Poor fits are instead observed with setting B, which also uses the LSM but PPE.st2. In this setting the cover of cereals (excluding *Secale*) is strongly underestimated as is (for the most part) the cover of *Secale*. Setting B overestimates the cover of grassland, *Alnus*, *Fagus* and *Pinus*; merely the reconstructed cover of *Picea* appears reasonable.

Also the performance of the model settings that use the GPM differs substantially. The overall poorest fit is found with setting C (GPM and PPE.MV2015). This setting produces too high reconstructed cover for cereals (excluding *Secale*), *Secale*, *Fagus* and *Picea* and too low cover for grassland, *Alnus* and *Pinus*. Model setting D (GPM and PPE.st2) performs better. It produces (partly) reasonable reconstructions for *Secale*, grassland and *Pinus* but arrives at too low cover for cereals (excluding *Secale*) and too high cover particularly for *Fagus*, less so for *Alnus* and *Picea*.



**Fig. 2** Reconstructed abundance of two taxa X and Y based on hypothetical pollen samples with 50 % pollen of X and 50 % pollen of Y. Samples taken from lakes and peatlands with 100–10,000 m diameter. X and Y have similar pollen productivity but pollen of X has a higher fall speed than that of Y. The *upper graphs* show the reconstructed plant cover; the lower graphs the K factors and their ratio. *Circles* denote results of REVEALS reconstructions using the GPM, *triangles* using the LSM

Model settings C and D were also calculated with the REVEALS programme of Shinya Sugita (settings C\* and D\*). The resulting mean cover values are similar to those found with the ‘REVEALSinR’ function. Apparently, the two implementations produce comparable results despite differences in e.g. the lake models. However, for setting C\* the Sugita programme produced much higher error ranges than ‘REVEALSinR’ (Fig. 4). The error estimates for herbs even well exceed the natural limits of percentage data. They are highest for cereals (excluding *Secale*) (618.5 %), which is the taxon with the smallest PPE (0.2). It appears that the use of such small PPEs is problematic in Sugita’s REVEALS programme. ‘REVEALSinR’ instead performs reasonably well also for taxa with small PPEs. Furthermore, only ‘REVEALSinR’ produces—as expected in percentage data—error estimates that are not symmetric.

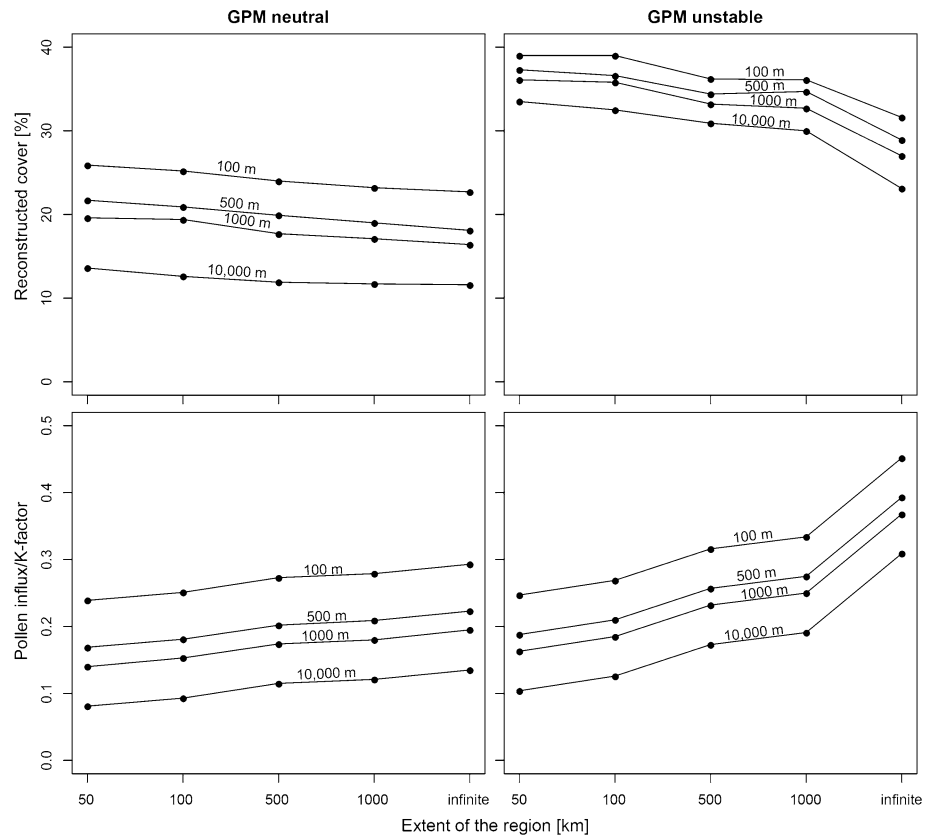
## Discussion

The considerable differences in model outcome and performance illustrate how important it is to select an appropriate dispersal model in REVEALS reconstructions. The two dispersal models that we tested differ substantially with respect to overall dispersal distances and the influence of pollen fall speeds. The pollen dispersal function enters the REVEALS reconstructions through the K-factor, which for each taxon represents predicted pollen influx at a site. The absolute value of K is not important in REVEALS, what matters is the difference between taxa. This difference is high in the GPM, where fall speed of pollen has a large effect on dispersal distances, but low in the LSM, where fall speed has only little effect. In other words, the GPM supposes a strong dispersal bias implying that independent of pollen productivity taxa with higher fall speed (such as *Fagus* and *Cerealia*) are under-represented in the pollen record of large basins compared with taxa with low fall speed (e.g. *Alnus* and grasses). REVEALS is designed to correct for this dispersal bias, but the choice of the dispersal model used for the correction can lead to large discrepancies in the reconstructions (Fig. 2).

## Dispersal model selection

Evidence shows that the LSM describes particle dispersal and deposition much better than the GPM. Upland pollen deposited in lakes or peatlands to a large degree arrives from some to many kilometres distance. Theuerkauf et al. (2013) showed that the LSM of Kuparinen et al. (2007) describes observed pollen deposition much better than the GPM; our data suggest the same (Table 3; Figs. 4, 5). Still, REVEALS applied with the GPM and PPE.st2 (settings D and D\*) also arrives at reasonable results for the Lake Tiefer See, except for *Cerealia* and *Fagus*. For both these taxa the poor fit could be attributed to unsuitable PPEs. The PPE for *Cerealia* in PPE.st2 derives from studies that include *Secale* in the analysis although *Secale* is wind-pollinated and emits far more pollen than the autogamous cereals *Avena*, *Hordeum* and *Triticum*. Because the PPE of *Cerealia* is too high, the resulting reconstructed cover is too low. Instead, in the case of *Fagus* the reconstructed cover is too high; suggesting that its PPE of 2.35 in the PPE.st2 dataset is too low. All studies from the lowlands of Central Europe indeed calculate higher PPEs between 5 and 15 (with grasses as reference and using the GPM; Sugita et al. 1999; Nielsen 2004; Theuerkauf et al. 2013; Matthias et al. 2012). With the *Fagus* PPE in setting D adjusted to 10, REVEALS produces a reasonable reconstruction also for *Fagus* (Fig. 5, dashed box).

**Fig. 3** The effect of the cut-off distance for the tail of the GPM in basins with 100, 500, 1,000 and 10,000 m diameter. The same model setting as in Fig. 2 with different limits to the extent of the region considered as pollen source area. The *upper graphs* show the reconstructed cover of Y; the lower graphs show K factors of Y. GPM for neutral conditions on the *left* and for unstable conditions on the *right*



**Table 4** Root mean square error of REVEALS based reconstructed plant cover. RMSE is calculated as distance between reconstructed cover and distance weighted plant abundance as recorded in written sources

RMSE	A	B	C	D	C*	D*
Cereals	7.5	20.5	20.9	15.1	17.4	16.0
Grassland	4.9	9.9	13.0	4.9	10.7	6.2
Secale	3.8	9.6	11.9	4.6	11.1	2.2
Pinus	5.4	2.3	4.1	0.4	2.6	0.3
Fagus	1.3	2.0	1.8	7.4	2.5	7.3
Alnus	0.4	3.5	2.0	0.6	1.7	1.1
Picea	3.5	0.4	5.8	2.8	6.1	2.7
Sum	26.8	48.2	59.5	35.9	52.1	35.7

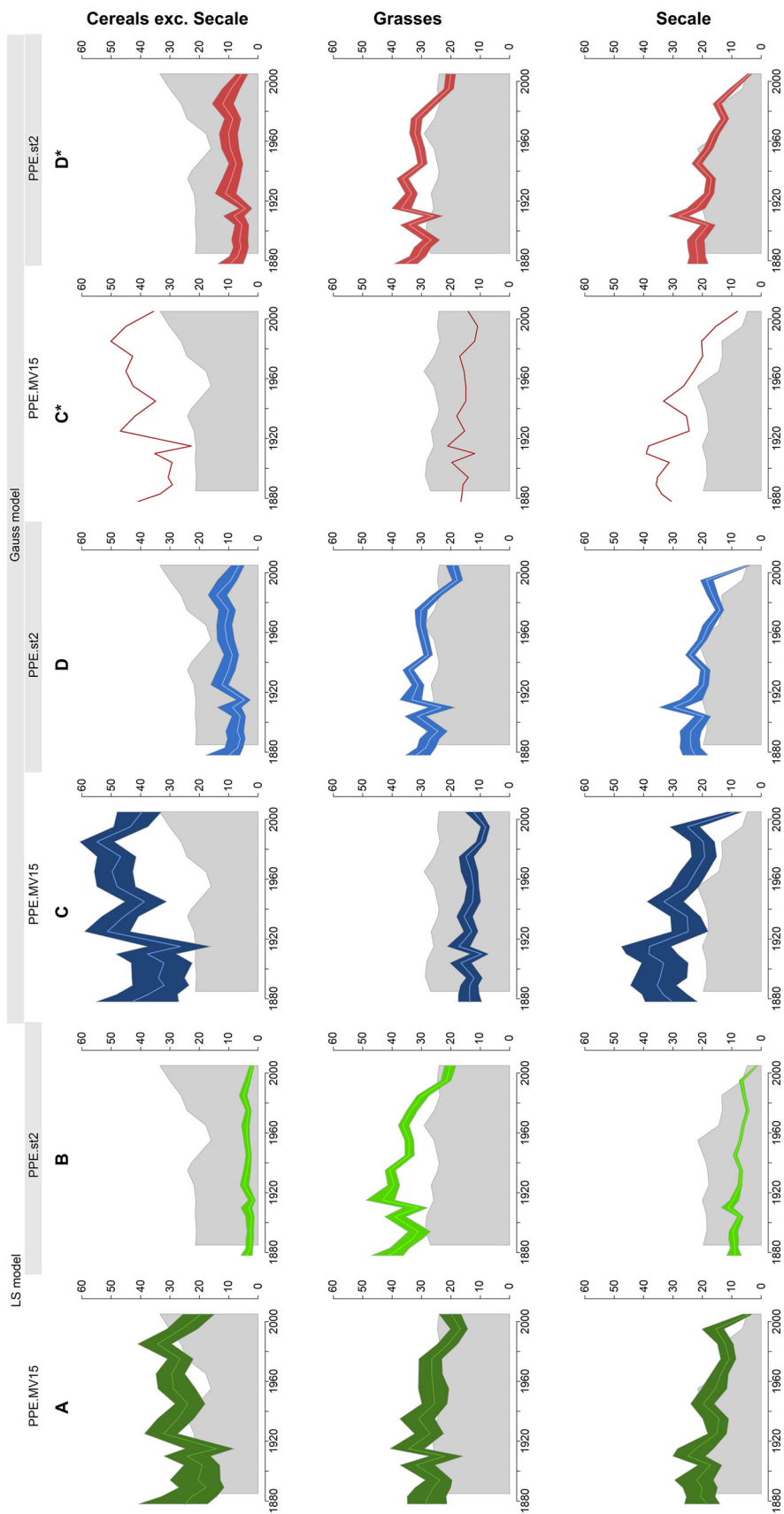
\* Repeated with Sugita's REVEALS programme (latest version: REVEALS.v4.2.2.Tallinn.wks.exe)

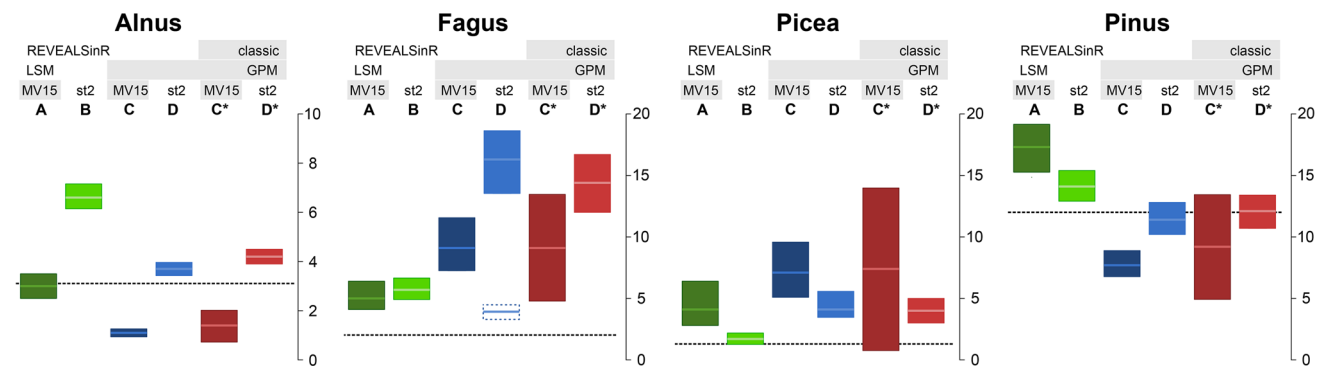
Apparently, REVEALS can produce satisfactory results with different dispersal models if PPEs are used that have been calculated in surface samples studies with the same underlying dispersal model and in basins of similar size. So, does the choice of dispersal model not matter after all? We argue that it does. First, we arrive at reasonable reconstructions with two very different sets of PPEs. Yet, obviously only one (if any) of these can truly be the set

that represents pollen productivity. PPEs are determined in studies that relate modern pollen data to modern plant abundances. Dispersal models are crucial in the calculation because they determine distance weighting (Theuerkauf et al. 2013); they provide an answer to the question how much of the pollen signal is arriving from nearby and how much from far away. This answer is not trivial, particularly if pollen fall speeds differ and have a strong effect on the resulting pollen signal, as in the case of the GPM (Fig. 2). Only if the dispersal model is appropriate, distance weighted abundances are correct and the resulting PPEs indeed represent the pollen productivity of the taxa involved. The GPM underestimates pollen dispersion in taxa with higher fall speed such as *Fagus* and *Secale*. As a result, distance weighted plant abundances are too low for these taxa, which the model compensates with a high PPE. Indeed, all studies from the lowlands of Central Europe produce a high PPE for *Fagus* when using the GPM (Sugita et al. 1999; Nielsen 2004; Matthias et al. 2012; Theuerkauf et al. 2013), although *Fagus* is commonly considered an intermediate pollen producer (Pohl 1937; Andersen 1970). To accommodate the expectation of a moderate PPE for *Fagus*, data have been discarded (Matthias et al. 2012) and averaged with low values from Switzerland (Mazier et al. 2012). Yet, to arrive at reasonable reconstructions of *Fagus* cover with the GPM, an



**Fig. 4** REVEALS based reconstructed cover of major crops (coloured shading) during the period 1880–2010 compared with the observed cover (gray shading). Upper and lower limits represent the 10 and 90 % percentile of 100 repeated model runs with settings A–D and standard error for setting D\*. The standard error for setting C\* is extremely high and not displayed





**Fig. 5** REVEALS based reconstructed cover (in colour bars) of major trees using the six model settings (Table 2) compared with modern cover (dashed line). Upper and lower limits of the boxes represent the 10 and 90 % percentiles of 100 repeated model runs

unreasonably high PPE is necessary. When using the GPM, PPEs are merely a correction parameter and do not truly represent (relative) pollen productivity, i.e. they are not pollen productivity estimates in the meaning of the word.

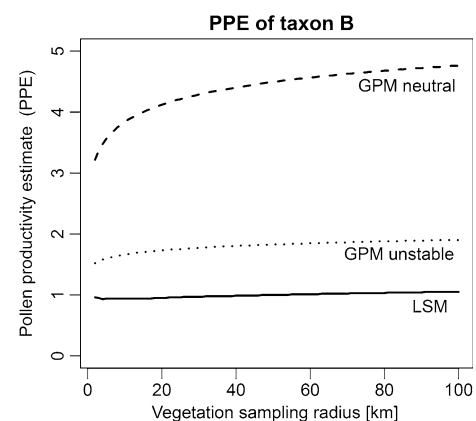
The problem is not only semantic, however. The use of an inappropriate dispersal model like the GPM will directly affect the REVEALS modelling results. Like R-values, PPEs calculated with an inappropriate dispersal model will differ between small and large basins and will not be universally applicable. Thus, PPEs that are calculated (using the GPM) from pollen-vegetation-relationships in small basins are not applicable to large basins (and vice versa). Moreover, it matters whether PPEs are calculated from the relationship between pollen and the vegetation in a short (e.g. 2 km) or a long distance around the basin (100 km), even in landscapes with homogenous vegetation cover. This effect is far more pronounced with the GPM than with the LSM (Fig. 6).

Another problem is related to the infinitely long tail of the GPM, particularly when light pollen types are concerned. In REVEALS studies the extent of a region is arbitrarily limited, mostly to 50 or 100 km (Mazier et al. 2012) and pollen modelled to arrive from more distant sources is neglected. However, with shorter cut-off distances progressively more pollen with low fall speed is neglected than pollen with high fall speed, affecting the model results. The effect is even more pronounced in the GPM adjusted for unstable conditions but largely absent in the LSM (Fig. 3).

### What is the region?

The dispersal models do not only affect the REVEALS calculations, but also matter in the interpretation of the results. REVEALS output is commonly interpreted as

with settings A–D and standard error for settings C\* and D\*. For Fagus, the dashed box indicates results with the PPE in setting D adjusted to 10 (see “Discussion” section)



**Fig. 6** PPEs calculated from a hypothetical pollen sample from a lake with a diameter of 2,000 m. Taxon X and Y each comprise 50 % of the surrounding, homogenous vegetation and 50 % of the pollen deposition in the lake. Pollen of X has a higher fall speed ( $v_g = 0.06 \text{ m s}^{-1}$ , similar to *Fagus*) than pollen of Y ( $v_g = 0.03 \text{ m s}^{-1}$ , similar to *Pinus*). The resulting PPE of X (with Y as the reference) is much higher when calculated with the GPM than with the LSM, because the GPM assumes that pollen of X is dispersed less far. With the GPM, the PPE of X also markedly increases with increasing sampling radius although vegetation composition is uniform

representing the regional vegetation composition—but how large is this region? Or, where does the pollen come from? There is no simple answer because pollen arrives from nearby as well as far away, with nearby sources contributing (much) more (Janssen 1966). Prentice and Webb (1986) suggested approximating the source area as the area outside the basin from which e.g. 80 % of total pollen deposition arrives. For large lakes and peatlands with 1,000 m diameter, the LSM predicts that the size of the 80 % source area is  $\sim 55 \text{ km}$  for all taxa, whether with high or low fall speed. In contrast, the conventional GPM for neutral conditions predicts a large difference in the 80 % source area of taxa with low ( $\sim 120 \text{ km}$ ) and taxa

with high fall speed (12 km; Table 1). Whereas the unrealistic GPM defies definition of a distinct source area, the realistic LSM offers a clear delineation.

The above calculations of the pollen source area assume that the vegetation cover of the region is homogenous. This is a central assumption in REVEALS modelling that is rarely met in reality. In the present study area, for example, vegetation follows a pattern that primarily reflects soil types (morainic sediments versus outwash plains). REVEALS-based vegetation reconstructions in such patchy landscapes may strongly differ from true abundances. The problem is most obvious in the disturbing effects that shore vegetation can have on the pollen record found in a lake. For example, high pollen values of *Alnus* in a lake may solely derive from a small fringe of *Alnus* trees around the lake (Janssen 1959). However, a REVEALS reconstruction would reconstruct *Alnus* as an important element of the regional vegetation.

Therefore, in situations where regional vegetation is expected to be patchy, approaches that do not rely on homogeneity are preferable to REVEALS. For a single site, multiple scenario approaches allow the detection of vegetation mosaics (Fyfe 2006; Bunting et al. 2008).

If pollen data are available from many sites, site to site differences in pollen deposition may be exploited to reconstruct patches, as it is done in the (extended) downscaling approach (Theuerkauf and Joosten 2009; Theuerkauf et al. 2014).

## REVEALSinR

We have implemented the REVEALSinR function in a way that allows for easy, rapid and automated application with full control of all parameters. REVEALSinR thus also provides a sandbox to test the effects and sensitivities of model assumptions and parameter settings, some of which we discuss above. Moreover, the robustness of reconstructions can be assessed by varying the parameter settings. For example, REVEALS is usually applied under the assumption that the pollen productivity of taxa is constant in time. In reality, however, pollen productivity is known to respond to changes in climate, stand density, soil conditions and land management. The effects are still poorly understood and have rarely been quantified (Feeser and Dörfler 2014; Theuerkauf et al. 2015). REVEALSinR enables running numerous PPE scenarios to establish variability and probabilities in reconstructions. Effects of error in pollen data can be assessed as well. REVEALSinR is able to deal with very small and large PPEs and in all cases produces reasonable, asymmetric error estimates. As mentioned above, the error estimates are only applicable in homogenous vegetation.

In its default settings REVEALSinR runs with the state-of-the-art LSM (and suitable PPEs), because this model is the most appropriate for describing regional pollen dispersal and deposition. Yet, like any model, this model also has its limitations. For example, in its current form the model is adjusted to atmospheric conditions that prevail in and above a pine forest. Furthermore, the model so far neglects diurnal changes in wind speed and turbulence. However, the model is flexible enough to include variations in these (and further) parameters.

## Conclusions

The choice of dispersal model matters in REVEALS reconstructions, much more than has hitherto been acknowledged. The commonly used GPM does not depict pollen dispersal well. If REVEALS is run with the GPM, the required PPEs do not represent pollen productivity of plant taxa, but rather a basin-specific correction factor. PPEs derived for one basin are not necessarily applicable to another and uncertainties ensue in reconstructions. Averaging PPEs over multiple studies will alleviate the inaccuracies to some extent, but does not address the core problem posed by an inappropriate dispersal model. We suggest that the GPM is replaced by the LSM both in REVEALS applications (including the LRA) as well as in the associated calculation of PPEs.

REVEALS produces mean regional plant abundances under the assumption of homogenous vegetation composition. In a patchy landscape, true vegetation composition may deviate considerably from the REVEALS reconstruction. To solve this discrepancy new approaches are needed. ‘REVEALSinR’ is only a first step in that direction.

Our R routine provides a tool that is open to further implementations. It offers a sandbox for testing model sensitivities and assessing consequences of parameter choices. The REVEALSinR function is part of the DISCOVER package (DIVERse Set of models for Quantitative pOLLEN-based VEgetation Reconstruction). Additional functions like MARCO POLO and extended downscaling are currently in the testing phase.

**Acknowledgments** We dedicate this paper to the memory of Roel Janssen and Herb Wright, whose knowledge and positive critical stance remain an inspiration. We thank Almut Mrotzek, Max Wenzel and Hans Joosten for fruitful discussions as well as John Birks and an anonymous reviewer for valuable comments on the manuscript. This study has utilized infrastructure of the Terrestrial Environmental Observatory (TERENO) of the Helmholtz Association and is a contribution to the Virtual Institute of Integrated Climate and Landscape Evolution Analysis—ICLEA—of the Helmholtz Association (VH-VI-415). The study was funded by Academy of Finland (AK).

## References

- Andersen ST (1970) The relative pollen productivity and pollen representativity of North European trees and correction factors for tree pollen spectra. *Danmarks Geol Undersøgelser Ser II* 96:1–99
- Bunting MJ, Twiddle CL, Middleton R (2008) Using models of pollen dispersal and deposition in hilly landscapes: some possible approaches. *Palaeogeogr Palaeoclimatol Palaeoecol* 259:77–91. doi:10.1016/j.palaeo.2007.03.051
- Davis MB (1963) On the theory of pollen analysis. *Am J Sci* 261:897–912
- Feeser I, Dörfler W (2014) The glade effect: vegetation openness and structure and their influences on arboreal pollen production and the reconstruction of anthropogenic forest opening. *Anthropocene* 8:92–100. doi:10.1016/j.ancene.2015.02.002
- Fyfe RM (2006) GIS and the application of a model of pollen deposition and dispersal: a new approach to testing landscape hypotheses using the POLLANDCAL models. *J Archaeol Sci* 33:483–493. doi:10.1016/j.jas.2005.09.005
- Giddings GD, Sackville Hamilton NR, Hayward MD (1997) The release of genetically modified grasses. Part 1: pollen dispersal to traps in *Lolium perenne*. *Theor Appl Genet* 94:1,000–1,006. doi:10.1007/s001220050507
- Gregory P (1945) The dispersion of air-borne spores. *Trans Br Mycol Soc* 28:26–72. doi:10.1016/S0007-1536(45)80041-4
- Hofmann F, Otto M, Wosniok W (2014) Maize pollen deposition in relation to distance from the nearest pollen source under common cultivation—results of 10 years of monitoring (2001 to 2010). *Environ Sci Eur* 26:24. doi:10.1186/s12302-014-0024-3
- Hultberg T, Gaillard M-J, Grundmann B, Lindbladh M (2015) Reconstruction of past landscape openness using the Landscape Reconstruction Algorithm (LRA) applied on three local pollen sites in a southern Swedish biodiversity hotspot. *Veget Hist Archaeobot* 24:253–266. doi:10.1007/s00334-014-0469-8
- Jackson ST, Lyford ME (1999) Pollen dispersal models in Quaternary plant ecology: assumptions, parameters, and prescriptions. *Bot Rev* 65:39–75
- Janssen CR (1959) *Alnus* as a disturbing factor in pollen diagrams. *Acta Bot Neerl* 8:55–58. doi:10.1111/j.1438-8677.1959.tb00005.x
- Janssen CR (1966) Recent pollen spectra from the deciduous and coniferous-deciduous forests of Northeastern Minnesota: a study in pollen dispersal. *Ecology* 47:804–825. doi:10.2307/1934267
- Klein E, Lavigne C, Foueillassar X et al (2003) Corn pollen dispersal: quasi-mechanistic models and field experiments. *Ecol Monogr* 73:131–150
- Kuparinen A (2006) Mechanistic models for wind dispersal. *Trends Plant Sci* 11:296–301. doi:10.1016/j.tplants.2006.04.006
- Kuparinen A, Markkanen T, Riikonen H, Vesala T (2007) Modeling air-mediated dispersal of spores, pollen and seeds in forested areas. *Ecol Model* 208:177–188. doi:10.1016/j.ecolmodel.2007.05.023
- Matthias I, Nielsen AB, Giesecke T (2012) Evaluating the effect of flowering age and forest structure on pollen productivity estimates. *Veget Hist Archaeobot* 21:471–484. doi:10.1007/s00334-012-0373-z
- Mazier F, Gaillard M-J, Kuneš P et al (2012) Testing the effect of site selection and parameter setting on REVEALS-model estimates of plant abundance using the Czech Quaternary Palynological Database. *Rev Palaeobot Palynol* 187:38–49. doi:10.1016/j.revpalbo.2012.07.017
- Mehl IK, Hjelle KL (2016) From deciduous forest to open landscape: application of new approaches to help understand cultural landscape development in western Norway. *Veget Hist Archaeobot* 25:153–176. doi:10.1007/s00334-015-0539-6
- Nathan R, Katul GG, Horn HS et al (2002) Mechanisms of long-distance dispersal of seeds by wind. *Nature* 418:409–413. doi:10.1038/nature00844
- Nathan R, Katul GG, Bohrer G et al (2011) Mechanistic models of seed dispersal by wind. *Theor Ecol* 4:113–132. doi:10.1007/s12080-011-0115-3
- Nielsen AB (2004) Modelling pollen sedimentation in Danish lakes at c. AD 1800: an attempt to validate the POLLSCAPE model. *J Biogeogr* 31:1,693–1,709. doi:10.1111/j.1365-2699.2004.01080.x
- Nurmiemi M, Tufto J (1998) Spatial models of pollen dispersal in the forage grass meadow fescue. *Evol Ecol* 12:487–502. doi:10.1023/A:1006529023036
- Parsons RW, Prentice IC (1981) Statistical approaches to R-values and the pollen—vegetation relationship. *Rev Palaeobot Palynol* 32:127–152
- Pohl F (1937) Die Pollenerzeugung der Windblüher. *Beih Bot Cent.bl* 56:365–470
- Prentice IC (1985) Pollen representation, source area, and basin size: toward a unified theory of pollen analysis. *Quat Res* 23:76–86. doi:10.1016/0033-5894(85)90073-0
- Prentice IC, Webb T (1986) Pollen percentages, tree abundances and the Fagerlind effect. *J Quat Sci* 1:35–43. doi:10.1002/jqs.3390010105
- R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.r-project.org/>
- Schmidt W (1918) Die Verbreitung von Samen und Blütenstaub durch die Luftbewegung. *Oesterr Bot Z* 67:313–328. doi:10.1007/BF02126080
- Soons MB, Heil GW, Nathan R, Katul GG (2004) Determinants of long-distance seed dispersal by wind in grasslands. *Ecology* 85:3,056–3,068. doi:10.1890/03-0522
- Sugita S (1993) Pollen dispersal model for an entire lake surface. *Quat Res* 39:239–244
- Sugita S (2007a) Theory of quantitative reconstruction of vegetation I: pollen from large sites REVEALS regional vegetation composition. *Holocene* 2:229–242
- Sugita S (2007b) Theory of quantitative reconstruction of vegetation II: all you need is LOVE. *Holocene* 2:243–258
- Sugita S, Gaillard M-J, Broström A (1999) Landscape openness and pollen records: a simulation approach. *Holocene* 9:409–421. doi:10.1191/095968399666429937
- Sutton OG (1947) The problem of diffusion in the lower atmosphere. *Q J R Meteorol Soc* 73:257–276
- Sutton OG (1953) *Micrometeorology: a study of physical processes in the lowest layers of the earth's atmosphere*. McGraw-Hill, New York
- Tackenberg O (2003) Modelling long-distance dispersal of plant diaspores by wind. *Ecol Monogr* 73:173–189
- Tauber H (1965) Differential pollen dispersion and the interpretation of pollen diagrams. *Danmarks Geol Undersøgelser Ser II* 89:1–69
- Theuerkauf M, Joosten H (2009) Substrate dependency of Lateglacial forests in north-east Germany: untangling vegetation patterns, ecological amplitudes and pollen dispersal in the past by downscaling regional pollen. *J Biogeogr* 36:942–953. doi:10.1111/j.1365-2699.2008.02047.x
- Theuerkauf M, Kuparinen A, Joosten H (2013) Pollen productivity estimates strongly depend on assumed pollen dispersal. *Holocene* 23:14–24. doi:10.1177/0959683612450194
- Theuerkauf M, Bos JAA, Jahns S et al (2014) *Corylus* expansion and persistent openness in the early Holocene vegetation of northern central Europe. *Quat Sci Rev* 90:183–198. doi:10.1016/j.quascirev.2014.03.002

- Theuerkauf M, Dräger N, Kienel U et al (2015) Effects of changes in land management practices on pollen productivity of open vegetation during the last century derived from varved lake sediments. *Holocene* 25:733–744. doi:[10.1177/0959683614567881](https://doi.org/10.1177/0959683614567881)
- Tufto J, Engen S, Hindar K (1997) Stochastic dispersal processes in plant populations. *Theor Popul Biol* 52:16–26
- von Post L (1918) Skogsträdpollen i sydsvenska torvmosslagerföljder. *Forhandlingar ved 16. Skandinaviske Naturforskeresmøte* 1916:433–465