Reliably Predicting Pollinator Abundance: Challenges of Calibrating Process-Based Ecological Models

Emma Gardner[1,2], Tom D. Breeze[2], Yann Clough[3], Henrik G. Smith[3], Katherine C. R. Baldock[4,5,6], Alistair Campbell[7], Mike Garratt[2], Mark A. K. Gillespie[8,9], William E. Kunin[8], Megan McKerchar[10], Jane Memmott[4], Simon G. Potts[2], Deepa Senapathi[2], Graham N. Stone[11], Felix Wäckers[12], Duncan B. Westbury[10], Andrew Wilby[12], Tom H. Oliver[1]

[1]School of Biological Sciences, University of Reading, Reading, UK

[2]Centre for Agri-Environmental Research, University of Reading, Reading, UK

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

[4]School of Biological Sciences, University of Bristol, Bristol, UK

[5]Cabot Institute, University of Bristol, Bristol, UK

[6]Department of Geographical and Environmental Sciences, Northumbria University, Newcastle upon Tyne, UK

[7]Laboratório de Entomologia, Embrapa Amazônia Oriental, Belém, Brazil

[8]School of Biology, University of Leeds, Leeds, UK

[9]Department of Environmental Sciences, Western Norway University of Applied Sciences, Sogndal, Norway

[10]School of Science and the Environment, University of Worcester, UK

[11]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

[12]Lancaster Environment Centre, Lancaster University, Lancaster, UK

Corresponding author:

Emma Gardner (e.l.gardner@reading.ac.uk)

School of Biological Sciences

Harborne Building

University of Reading

Whiteknights

Reading RG6 6AS

Running headline: Reliably Predicting Pollinator Abundance

Abstract

1. Pollination is a key ecosystem service for global agriculture but evidence of pollinator population declines is growing. Reliable spatial modelling of pollinator abundance is essential if we are to identify areas at risk of pollination service deficit and effectively target resources to support pollinator populations. Many models exist which predict pollinator abundance but few have been calibrated against observational data from multiple habitats to ensure their predictions are accurate.

2. We selected the most advanced process-based pollinator abundance model available and calibrated it for bumblebees and solitary bees using survey data collected at 239 sites across Great Britain. We compared three versions of the model: one parameterised using estimates based on expert opinion, one where the parameters are calibrated using a purely data-driven approach and one where we allow the expert opinion estimates to inform the calibration process.

3. All three model versions showed significant agreement with the survey data, demonstrating this model's potential to reliably map pollinator abundance. However, there were significant differences between the nesting/floral attractiveness scores obtained by the two calibration methods and from the original expert opinion scores.

4. Our results highlight a key universal challenge of calibrating spatially-explicit, process-based ecological models. Notably, the desire to reliably represent complex ecological processes in finely mapped landscapes necessarily generates a large number of parameters, which are challenging to calibrate with ecological and geographical data that is often noisy, biased, asynchronous and sometimes inaccurate. Purely data-driven calibration can therefore result in unrealistic parameter values, despite appearing to improve model-data agreement over initial expert opinion estimates. We therefore advocate a combined approach where data-driven calibration and expert opinion are integrated into an iterative Delphi-like process, which simultaneously combines model calibration and credibility assessment. This may provide the best opportunity to obtain realistic parameter estimates and reliable model predictions for ecological systems with expert knowledge gaps and patchy ecological data.

Keywords— calibration, credibility assessment, Delphi panels, ecosystem services, pollinators, process-based models, validation

1 Introduction

Pollination is a key ecosystem service underpinning the reproduction of many flowering plants, including many crops. Pollinators enhance production in $\sim 75\%$ of globally significant crops, adding $>$ \$235bn p.a. of productivity and substantially increasing the nutritional security of people the world over (Smith et al., 2015; Breeze et al., 2016). However, pollinator populations are under increasing pressure from landscape simplification (Kennedy et al., 2013), agrochemical use (Rundlöf et al., 2015; Woodcock et al., 2017) and climate change (Kerr et al., 2015), and there is growing evidence of instability in pollinator-dependent crop yields (Garibaldi et al., 2011; Garratt et al., 2014). Unless addressed, these pressures are expected to cause significant declines in global pollinator diversity in the coming decades (Rasmont et al., 2015; Balfour et al., 2018), threatening global food

security.

To date, very few countries have sufficient data to monitor pollinator abundance (O'Connor et al., 2019) or diversity (Carvalheiro et al., 2013, Kerr et al., 2015; Powney et al., 2019) and so cannot reliably identify areas suffering declines or at risk of sub-optimal pollination services (Garibaldi et al., 2011). Although field monitoring of national scale trends in pollinators and pollination services is both scientifically and economically viable (Breeze et al., 2019; O'Connor et al., 2019), it will take several years to build up such databases. Until then, additional approaches are needed to help target resources to support pollinator populations.

Spatial modelling of pollinator populations can support decision-making and is essential to predict the effects of future land-use change on pollinator populations. The most simplistic spatial models of pollination are purely based on crop forage distance from semi-natural habitat (Priess et al., 2007). Other studies assign habitat quality scores to all habitat types in the landscape (Schulp et al., 2014; Nogué et al., 2016), but this does not capture the fact that pollinators may use different habitats for different resources. The more sophisticated InVEST pollinator model, developed by Lonsdorf et al. (2009), assigns a separate nesting and flowering quality score to each habitat for different taxa, accounting for flight distances. This model and adaptations of it have already been used to infer spatially explicit current (Koh et al., 2016; Zhao et al., 2019) and future trends in pollinators/pollination (Chaplin-Kramer et al., 2019) and estimate pollinator natural capital (Ricketts and Lonsdorf, 2013).

More recent studies have refined this process-based InVEST model further by assuming that pollinators are optimal foragers (Olsson et al., 2015), accounting for temporal variation in floral resources and using expert-derived floral attractiveness scores (Häussler et al., 2017). If models are to be capable of predicting the impact of future land-use change on pollinators and reliably informing conservation management, such sophisticated and realistic simulation of pollinator requirements and resource use is essential. The most advanced social bee models currently available — BEEHAVE and Bumble-BEEHAVE (Becher et al., 2014; Becher et al., 2018) — adopt an agent-based approach, simulating the behaviour of individual bees. However, such agent-based modelling is computationally intensive, such that process-based models remain the most viable option for predicting pollinator visitation across large spatial scales, while still accounting for fine-grained differences in land-use.

Structural realism alone is not sufficient, however, and any model used to inform land management and policy must also be validated against observational data to ensure its predictions reflect current observed reality. Several studies have compared predictions from process-based pollinator models with field data (e.g. Lonsdorf et al., 2009; Kennedy et al., 2013; Ricketts and Lonsdorf, 2013; Groff et al., 2016; Nicholson et al., 2019), but these have primarily focused on predicting pollinator abundance within specific crops. Such models have yet to be validated more widely using pollinator abundance measurements in both crop and non-crop landcovers.

Here, we take the most advanced process-based pollinator abundance model available (ScaLE-poll; Häussler et al., 2017), which simulates both solitary and social bees (the main UK pollinators of crops and wild flowers), and we compare its predictions to abundance data collected at 239 sites across Great Britain, including crop, non-crop and urban sites. Our aim is to identify an optimum set of parameters for the model that produce the best agreement with the observed survey data and enable the model to be used with confidence to predict the consequences of land-use change on UK pollinator populations and pollination service. We first parameterise the model using nesting and floral attractiveness scores derived from expert opinion and assess the level of model-data

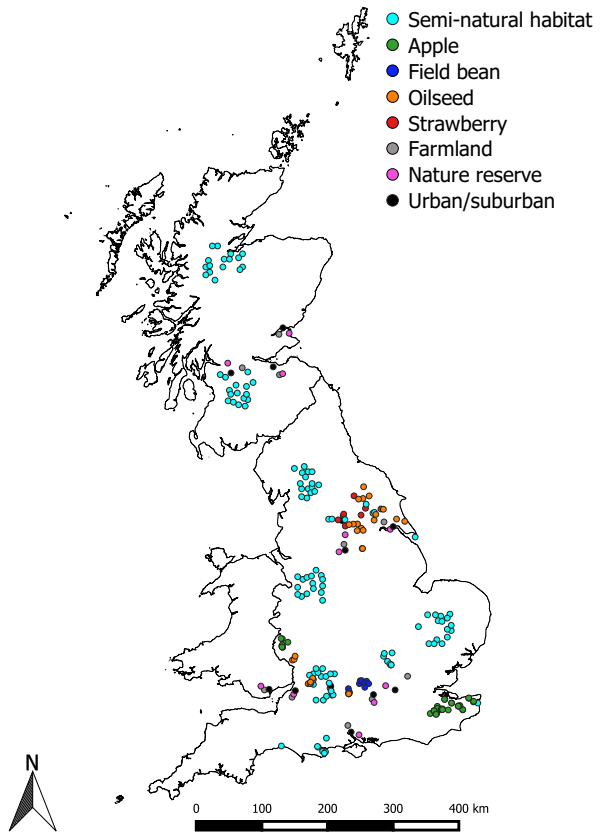

Figure 1: Locations of survey sites colour coded by type of survey site (see Table S1 for type definitions).

agreement. We then use the observed abundance data to calibrate these nesting and floral attractiveness scores and improve the model-data agreement, using an Approximate Bayesian Computation-like approach. We test two calibration methods: a free data-driven calibration and an expert-informed prioritised calibration. We discuss the implications of these three different model parameterisations, the realism of their derived parameter values and their implications for reliably modelling pollination service at large spatial scales.

2 Materials and Methods

2.1 Pollinator Abundance Data

We collated transect data from surveys conducted between 2011 and 2016 at 239 sites across Great Britain (Fig. 1; Table S1), including 84 crop sites, 12 urban sites and 143 non-crop sites (i.e. nature reserves and semi-natural habitat). Number of surveys per site ranged from 1–14 (mean $= 4.5 \pm 0.1$ surveys per site).

For each survey, we sum up the total number of individuals observed within each of four guilds, which we can then compare to the model predictions, controlling for total transect length and survey date. The guilds are ground nesting bumblebees (GNBB), tree nesting bumblebees (TNBB), ground nesting solitary bees (GNSB) and cavity nesting solitary bees (CNSB), with species allocated to guilds following the nesting preferences given in Falk (2015). Where observations were not recorded to species level but instead recorded as '*Bombus* unknown' or 'solitary unknown', we divide these unknown individuals between the nesting guilds according to the proportions

of known individuals assigned to each guild on that particular survey. In practice, unknown *Bombus* and unknown solitary bees were predominantly assigned to ground nesting guilds due to observations of ground nesting species significantly outnumbering other guilds.

2.2 Model Description

ScaLE-poll is a process-based model that predicts spatially explicit abundance and flower visitation rates by wild central-place-foraging pollinators in a given landscape. It accounts for population growth over time, allows different dispersal distances for workers and reproductives, includes preferential use of more rewarding floral and nesting resources, and can incorporate fine-scale edge features in the landscape. We summarise the model below. For a detailed description of the model see Häussler et al. (2017).

The model requires a rasterised landcover map detailing the landcover class (e.g. cereal, woodland, etc.) of each pixel, as well rasters containing the area within each pixel that is covered by specific edge features (e.g. hedgerows, flower margins). Each landclass is scored according to the amount of floral cover it provides during each season (spring, summer and autumn), the attractiveness of those floral resources to each pollinator guild (floral attractiveness) and the attractiveness of the nesting opportunities the landclass provides to each pollinator guild (nesting attractiveness). For each guild, the model then generates a nesting resources map (i.e. nesting attractiveness score for each pixel multiplied by the pixel area and maximum nest density input into the model), plus floral resource maps for each season (i.e. floral attractiveness multiplied by seasonal floral cover score for each pixel).

Nests are then randomly allocated across the landscape, with the number of nests in a pixel drawn randomly from a Poisson distribution around the expected number predicted by the nesting resources map. For each nest, the model uses the foraging distance of the pollinator to calculate the resources gathered by the nest from its surroundings, which in turn determines how many workers (if social) and new queens the nest produces using the input growth parameters for that pollinator. New queens then disperse according to the dispersal kernel of the pollinator. In any given pixel, the number of new queens that survive to the following year is limited by the expected number of nests in that pixel according to the nesting resources map.

The model outputs visitation rate to each pixel in each season (based on the amount of time pollinators from all nests spend foraging in each pixel). Solitary bees are assumed to be active only during one season, with new bees produced at the end of this season. Social bees (e.g. bumblebees) are assumed to be active in all three seasons, with queens foraging during season 1, workers foraging during seasons 2 and 3, and new queens produced at the end of season 3.

2.3 Model Inputs

2.3.1 Landcover/Edgecover Rasters

Landcover rasters are generated from the CEH Land Cover Map 2015 (LCM2015) with Ordnance Survey orchard polygons added on top of this. Where a land parcel is classed as 'Arable and Horticulture' in LCM2015, we obtain crop information for the year 2016 from rural payments agency databases.

For each landcover raster, we also generate edgecover rasters for six edge features (ditches, fallow field margins, grassy field margins, flower-rich margins, hedgerows and woodland edges) using information from rural payments agency databases and the CEH Woody Linear Features Database (Scholefield et al., 2016). See Supplementary Material for full details of landcover/edgecover raster generation.

For each survey site, we generate 10x10km landcover/edgecover rasters with 10x10m pixels centred on the survey site, which are used to obtain model predictions for calibration and validation. To obtain upscaled calibrated model predictions for Great Britain, we also generate 512 35x35km landcover/edgecover rasters with 10x10m pixels, which cover the entire geographical area with a 5km overlap between rasters (later removed from output rasters to eliminate edge effects).

2.3.2 Expert Opinion Data: Floral cover, floral attractiveness and nesting attractiveness

Ten UK pollinator experts were asked to score 35 common European landclasses (Table S2) for abundance and duration of floral resources per season (later multiplied to obtain floral cover). They were also asked to assign floral and nesting attractiveness scores to each landclass for the pollinator guilds they had experience of. Scores were collected on a six point scale, along with corresponding 'certainty scores'. We then calculated the mean scores across all experts and their variance, weighted by the experts' certainty scores. See Supplementary Material for full details.

2.3.3 Literature Data: Maximum nest density, foraging distances, dispersal distances and growth parameters

We use the maximum nest density, foraging distance, dispersal distance and growth parameters supplied with the ScaLE-poll model for bumblebees and solitary bees and used in Häussler et al. (2017) (Table 1). For simplicity, we assume both bumblebee nesting guilds have the same values for these parameters. Similarly, we assume both solitary bee nesting guilds have the same values for these parameters. This is unlikely to be true. However, the identical maximum nest density assumption is unimportant for our results, since we never compare the relative abundance of guilds and are concerned only with calibrating relative attractiveness of landclasses within guilds. Similarly, we consider the uniform foraging and dispersal distances for bumblebees and solitary bees an appropriate simplification, since foraging and dispersal distances are poorly known and vary between species (even within guilds) and we compare our model predictions to observed guild totals of varying species composition.

2.4 Comparison of Model Predictions with Pollinator Abundance Data

To obtain a model prediction for a given survey site, we input the site's 10x10km landcover/edgecover rasters and calculate the predicted spring visitation rate per m^2 within the survey area (V_1) by summing up the season 1 visitation rate to all pixels inside the survey area and dividing by the total survey area. We compare this to the observed number of bees on each survey (N_{obs}) by fitting the model:

$$\log\left(\frac{N_{obs}+1}{L}\right) = \beta \log V_1 + \gamma \log W + \begin{pmatrix} \alpha_{2011} \\ \vdots \\ \alpha_{2016} \end{pmatrix} Y \quad (1)$$

where L is the total transect length (i.e. we implicitly assume bees are detected within some unknown width either side of the transect which is constant across sites), W is the week of the year that the survey was carried out, Y is the year the survey was carried out and we fit to $(N_{obs}+1)$ to avoid taking the logarithm of zero when no pollinators were recorded. The co-variable W allows us to account for the fact that pollinator population size changes during the survey season, e.g. as bumblebee nests produce workers over time and solitary bees' active periods pass. The co-variable Y allows us to account for the fact that pollinator abundance nationally shows between-year variability due to year-year variation in weather suitability impacting pollinator growth directly (e.g. through poor weather reducing foraging time) and indirectly (e.g. by reducing floral cover).

Although the survey data represent counts, we fit the linear model assuming a Gaussian error distribution rather than Poisson, because the count data are over-dispersed with variance much larger than the mean. We choose a Gaussian error distribution with logged variables rather than any other method to deal with over-dispersion, such as quasi-Poisson distribution, because this approach produces the smallest and most uniform residuals across the data range.

We fit the linear model using R version 3.5.1 (R Core Team, 2018). A positive value of β that is significantly different from zero indicates significant model-data agreement.

2.5 Sensitivity Analysis

We conduct a sensitivity analysis to determine how sensitive the model-data agreement is to changes in the input nesting and floral attractiveness scores. For each guild, we calculate the change (Δ) in model-data agreement slope (β; obtained from fitting Equation 1) when each attractiveness parameter is adjusted by $\pm 50\%$. This is done by running the ScaLE-poll model twice for each attractiveness parameter — once with that parameter increased by 50% and once with it decreased by 50%, whilst holding all other parameters constant at their original expert opinion values. For attractiveness parameters that are zero, we vary the parameter by $\pm 50\%$ around a value of 0.1. For each attractiveness parameter, we obtain model predictions across all the survey sites for these two scenarios (parameter $\pm 50\%$) and fit Equation 1 to obtain the model-data agreement slope in each scenario (β_+ and β_-). We then calculate the percentage change in the model-data agreement slope as:

$$\Delta = 100\frac{|\beta_+ - \beta_-|}{\beta} \quad (2)$$

where β is the model-data agreement slope when all attractiveness parameters are set to their original expert opinion values.

We calculate the uncertainty in Δ by propagating the standard errors on the individual slopes (α_{β_+}, α_{β_-} and α_β), following Hughes and Hase (2010), as:

$$\alpha_{\Delta} = \Delta \left[\left(\frac{\alpha_{\beta_+}^2 + \alpha_{\beta_-}^2}{\beta_+ - \beta_-} \right)^2 + \left(\frac{\alpha_{\beta}}{\beta} \right)^2 \right]^{1/2} \quad (3)$$

2.6 Model Calibration

We separate the survey sites into 120 calibration sites and 119 validation sites, using stratified random sampling to ensure both subsets contain equal proportions of crop/non-crop sites and zero/non-zero surveys per guild. The validation sites are not used for calibration but reserved for assessing improvement in model-data agreement.

For each guild, we focus on calibrating the nesting and floral attractiveness scores for each landclass, excluding the landclasses buckwheat (which does not occur in any of our survey site rasters) and 'unsuitable' (which is used for water/bare rock e.t.c. and is fixed at zero attractiveness). We keep the floral cover scores fixed at their original expert opinion values, to allow us to decouple the guilds and calibrate each guild separately, and all other parameters remain fixed at their literature values.

We test two different methods of calibration. Method 1 involves searching the parameter space of all eligible parameters simultaneously (free data-driven calibration). Method 2 (expert-informed prioritised calibration) involves first prioritising parameters for calibration according to the results of the sensitivity analysis and searching the parameter space of parameters which the model is most sensitive to first, while less sensitive parameters remain fixed at their original expert opinion values. The parameter space of these less sensitive parameters is only searched once the more sensitive parameters have been calibrated. We define three sensitivity thresholds for Method 2: parameters which produce $\Delta \geq 5\%$ are calibrated first, followed by parameters which produce $5\% > \Delta \geq 0.5\%$, with the remaining parameters producing $\Delta < 0.5\%$ calibrated last.

The calibration process itself follows an Approximate Bayesian Computation-like approach (Fearnhead and Prangle, 2012) and involves running ScaLE-poll 1000 times across all of our calibration sites. Each run uses a unique set of attractiveness parameters where each eligible attractiveness parameter is assigned a random value drawn from a uniform distribution between the allowable limits for that parameter, while any ineligible attractiveness parameters remain fixed at their original expert opinion scores. For each run, we fit Equation 1 to assess the model-data agreement between the calibration site survey data and model predictions and select the 100 runs which produce (β,R^2) closest to (1,1). We then calculate the density distributions of the eligible parameters across these 100 best runs. While the density distributions of the eligible parameters across all 1000 runs are flat, the density distributions corresponding to the 100 best runs should be biased towards parameter values that produce the best fit to the data and show a peak around this value. If the full width half maximum (FWHM) of the density distribution peak is $\leq 60\%$ of the parameter's allowable range, then we assume the parameter has been sufficiently constrained and we define the parameter's calibrated score as the score that corresponds to the density distribution peak. The FWHM limit of $\leq 60\%$ was set after careful consideration of the density distribution widths the calibration process and data quality are capable of producing.

Typically only a few ($\sim 1-5$) parameters will be constrained from analysing a single batch of 1000 runs, due to the large number of parameters being varied simultaneously broadening the density distributions of individual parameters. After a single batch of 1000 runs, any calibrated parameters are set to their calibrated values and the process is repeated until all parameters have been calibrated or the remaining parameters to do not yield FWHM

$\leq 60\%$ due to their minimal leverage on the model-data agreement.

2.7 Predicting Visitation Rates across Great Britain

We generate two model predictions per guild for spring visitation rate across Great Britain: one prediction using the calibrated attractiveness scores obtained using Method 1 (V_{cal}) and one using the original expert opinion attractiveness scores (V_{exp}). We compare the ratio of these two predictions by calculating V_{cal}/V_{exp} to identify regions of the country where these two predictions differ.

We quantify the uncertainty on V_{cal} by running 100 simulations, where the score of each attractiveness parameter is randomly selected from the density distribution of that calibrated parameter. The uncertainty on the model prediction in each pixel is then represented by the standard deviation of these 100 simulations.

We quantify the uncertainty on V_{exp} by running 100 simulations, where the score of each attractiveness parameter is randomly selected from a beta distribution ($B(a, b)$) with mean ($\mu = a/(a + b)$) and variance ($\sigma^2 = \mu(1 - \mu)/(a + b + 1)$) equal to the mean and variance of the expert opinion score for that parameter. Since $B(a, b)$ is only defined on the interval (0,1), we rescale the floral attractiveness parameter means (originally scored from 0–20) and variances onto the interval (0,1), draw randomly from the appropriate beta distribution and multiply the randomly selected scores by 20 to return them to the appropriate scale.

The uncertainty on the ratio V_{cal}/V_{exp} is taken as the standard deviation of the ratios calculated from dividing one V_{cal} simulation by the corresponding V_{exp} simulation. We assess the significance of the ratio V_{cal}/V_{exp} in each pixel by calculating the number of standard deviations the ratio is away from a ratio of 1:1 in that pixel, i.e. by expressing $(V_{cal}/V_{exp} - 1)$ in units of the standard deviation of V_{cal}/V_{exp} within that pixel. Pixels with $(V_{cal}/V_{exp} - 1) \geq 3$ standard deviations are considered to show a significant difference between the calibrated and expert opinion model predictions.

3 Results

3.1 Initial Model-Data Comparison

All four guilds show significant model-data agreement (i.e. statistically significant $\beta > 0$) between the initial model predictions for each survey site, calculated using the expert opinion attractiveness scores, and the observed survey data (Table 2). However, the agreement is non-linear with $\beta \ll 1$ for all guilds, implying a doubling of predicted visitation is not reflected by a doubling in observed abundance. R^2 values for the fits range from 0.285–0.467, with ground nesting guilds showing lower R^2 values than the other guilds.

3.2 Sensitivity Analysis

For each guild, only a small number of parameters produce a significant change in model-data agreement slope when adjusted (Fig. 2). For GNBB, only the nesting attractiveness of unimproved permanent grassland and cereal and the floral attractiveness of coniferous woodland, unimproved permanent grassland, cereal and oilseed produce a percentage change in slope with uncertainty bounds that do not overlap zero. This is due to the large

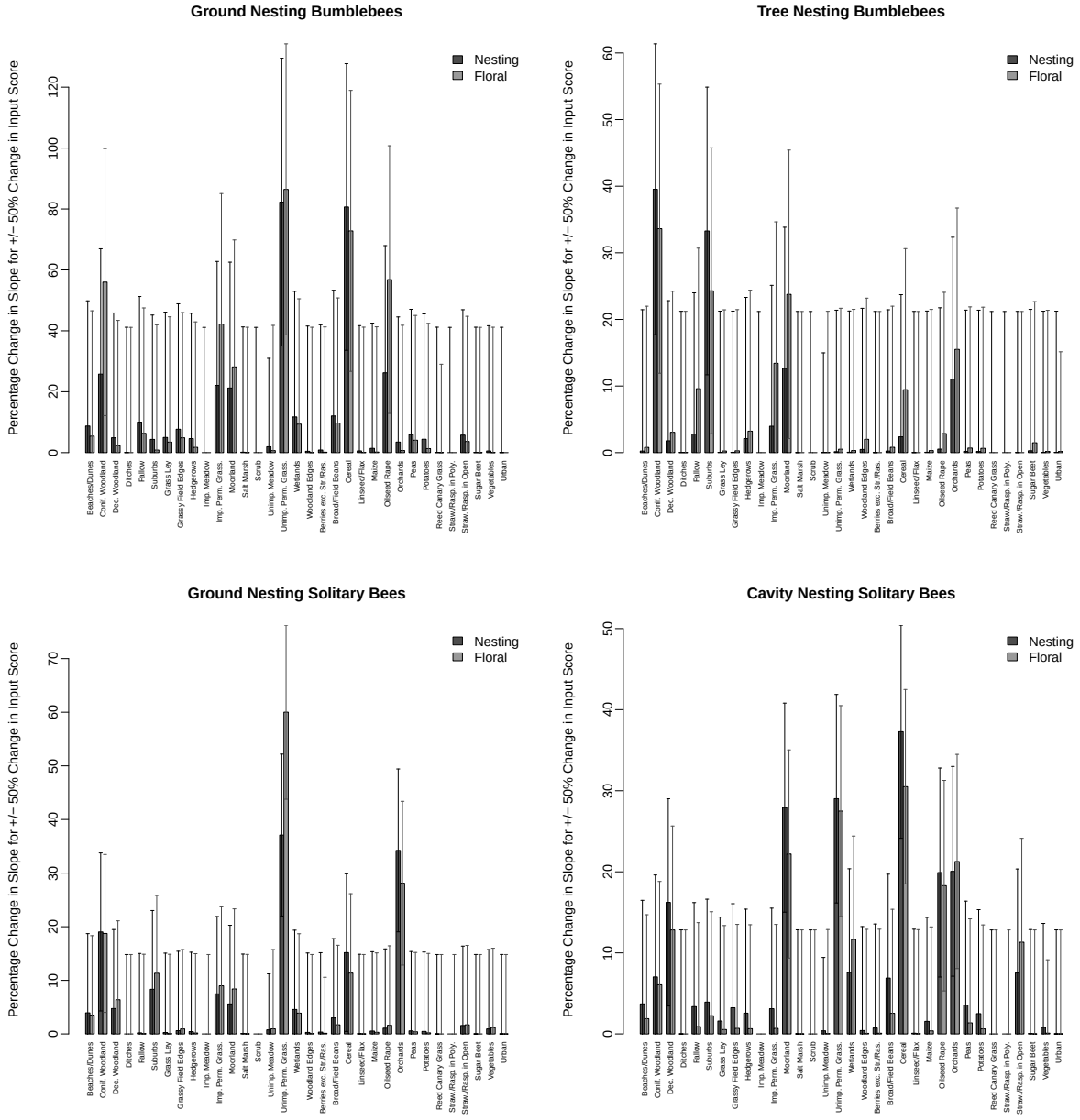

Figure 2: Sensitivity analysis results for each guild. Bar heights show percentage change in model-data agreement slope when landclass attractiveness scores are adjusted by ±50 %. Errorbars show propagated uncertainty using Eqn.3. See Table S2 for abbreviations.

scatter in the model-data agreement producing $\sim 10\%$ uncertainty on individual model-data slopes and the small geographic area covered by many landclasses (e.g. vegetables).

The model-data agreement sensitivity is influenced by landscape composition and data collection location, as well as incorporating expert opinion as to which landclasses should be important to each guild. It is most sensitive to: 1. landclasses that occur within many survey areas (e.g. orchards; Fig. 2 GNSB panel), 2. landclasses that occur close to survey areas and/or cover a large area within the surrounding landscape (e.g. cereal; Fig. 2 GNBB panel), 3. landclasses that have floral attractiveness scores similar to adjacent landclasses (such that flipping the score above/below that of an adjacent landclass produces a big change in where bees are foraging), and 4. landclasses that have high floral/nesting expert opinion attractiveness scores, since +/-50% of a high score results in a bigger absolute change in input attractiveness score than +/-50% of a low score (e.g. suburbs, Fig. 2 TNBB panel).

3.3 Model Calibration

3.3.1 Improvement in Model-Data Agreement

All four guilds show β closer to 1 and improved R^2 values after calibration (except Method 2 for TNBB), with R^2 values for all guilds now ranging from 0.358–0.482 (Table 2). Calibration Method 1 generally produces a slightly higher R^2 than Method 2, but there is typically no significant difference between the model-data agreement slopes obtained by the two methods, with TNBB the only guild for which the standard errors on the two slopes do not overlap.

The results in Table 2 represent the model-data fit agreement using all survey sites. Fig. 3 (and the corresponding Fig. S1, S2 and S3 in the Supplementary Material) show the improvement in β and R^2 as successive batches of parameters are calibrated for the calibration and validation sites separately. The validation sites, which were not used to calibrate the model, generally show a similar improvement in model-data agreement to the calibration sites, with the exception of TNBB R^2 using Method 1 (Fig. S1). In this case, the calibration subset began with a lower R^2 than the validation subset and selecting for (β, R^2) close to $(1, 1)$ in the calibration subset produces a slight reduction in R^2 for the validation subset.

Across all four guilds, the biggest improvements in model-data agreement occur at the beginning of both calibration processes, when the most influential parameters are calibrated (despite these not being forcibly prioritised by Method 1). GNBB show no further significant improvement in model-data agreement slope via Method 1 after the first six batches of parameters have been calibrated (18 out of 66 parameters). The optimal β value is typically achieved faster via Method 2 than Method 1. Method 2's prioritising of slope-influencing parameters means that improvements in R^2 often take longer to achieve than improvements in β, whereas β and R^2 improve at roughly the same rate using Method 1 (Fig. 3). For TNBB, this prioritisation of slope-influencing parameters actually results in an overall reduction in R^2 by the end of the Method 2 calibration process (Fig. S1). This may be due to TNBB being most restricted by the Method 2 calibration process, with just 11 parameters falling below the first sensitivity threshold for calibration, compared to 23, 15 and 19 parameters for GNBB, GNSB and CNSB guilds, respectively.

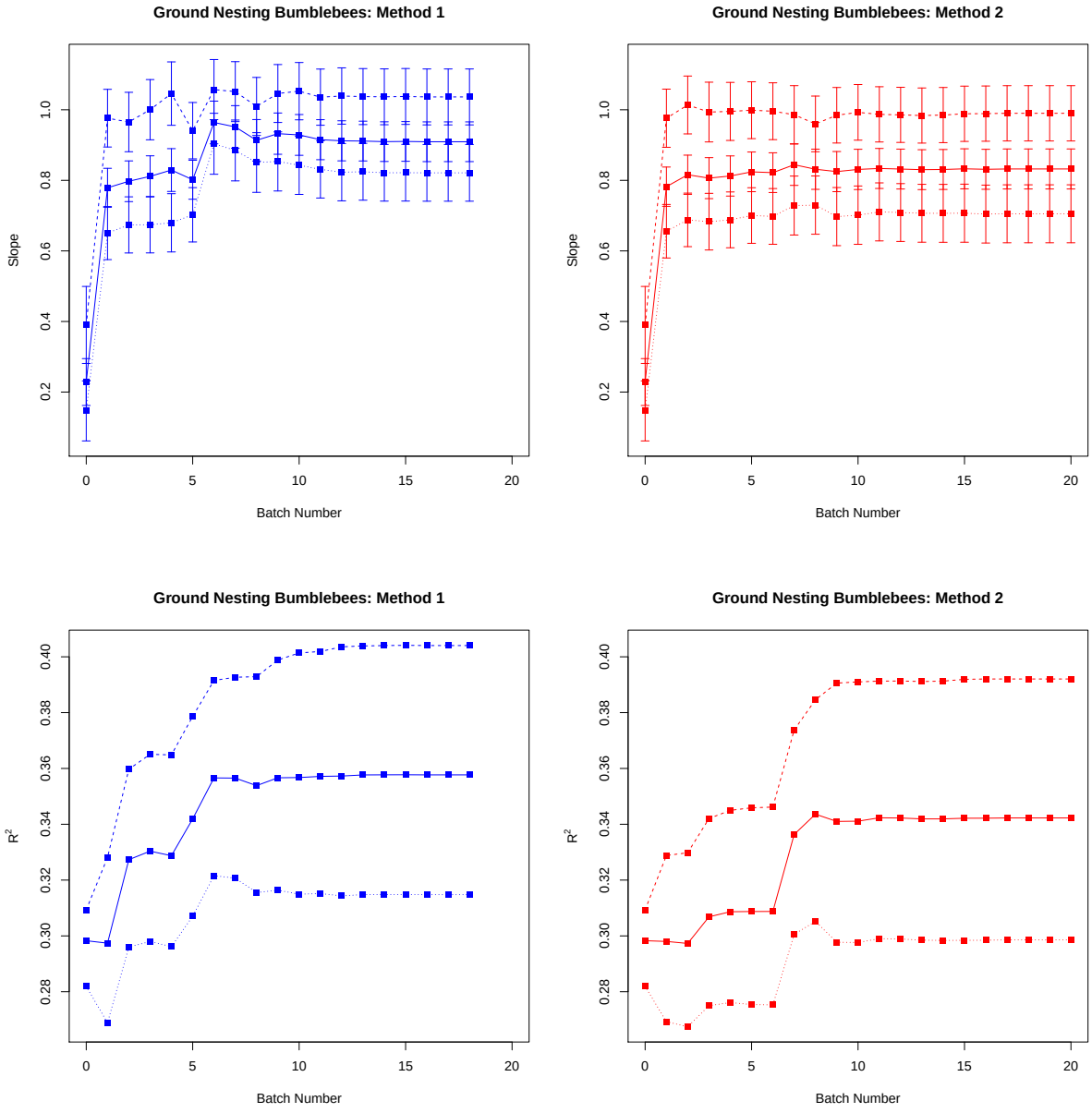

Figure 3: Change in model-data agreement slope and R^2 after each successive round of calibration using Methods 1 and 2 for Ground Nesting Bumblebees. Solid line shows results from fitting all survey sites, dashed and dotted lines show results for calibration and validation sites, respectively. Errorbars show slope standard error. Fig. S1, S2 and S3 in the Supplementary Material show corresponding plots for the other pollinator guilds.

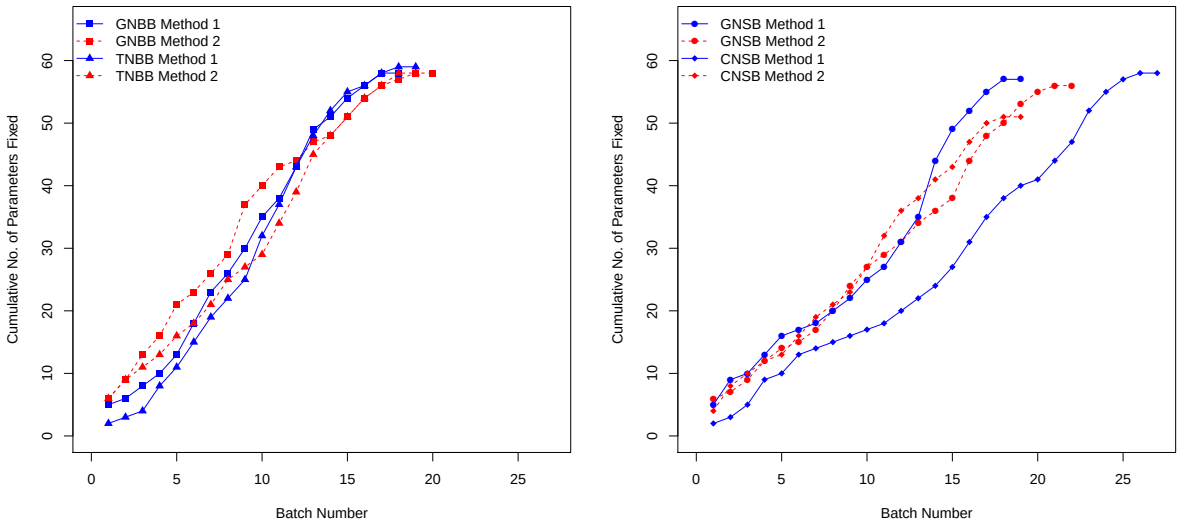

Figure 4: Comparison of rates at which parameters are fixed by successive rounds of calibration using Methods 1 and 2 for each guild.

3.3.2 Calibration Rates

For GNBB and TNBB (Fig. 4, left panel) the cumulative number of parameters fixed per patch is initially higher using Method 2 but drops below Method 1 for later batches, such that both calibration methods require the same total number of batches. However, the Method 2 calibration rate remains higher than Method 1 while the calibration process is still producing significant improvements in model-data agreement (cf. Fig. 3 and S1), only dropping below Method 1 after the overall change in model-data agreement becomes negligible. This suggests adopting Method 2 may be advantageous for these guilds, although TNBB may benefit from lower Δ thresholds to avoid the over-prioritisation of slope improvement at the expense of improvements in R^2.

For GNSB, the cumulative number of parameters fixed by Method 2 is always comparable to or less than Method 1, such that adopting Method 2 offers no advantage for this guild (circles, Fig. 4). In contrast, for CNSB, the calibration rate by Method 2 is always substantially higher than using Method 1 (diamonds, Fig. 4). Method 2 fixes a lower total number of parameters for CNSB than Method 1 (51 versus 58, respectively). However, Fig. S3 shows that significant improvements in model-data agreement ceased around batch 16 for both methods for this guild, at which point Method 2 had calibrated a greater number of parameters than Method 1.

3.3.3 Calibrated Attractiveness Scores versus Original Expert Opinion Scores

For individual attractiveness parameters, there can be large differences between the original expert opinion scores and the calibrated scores. For bumblebees, where both calibrated *nesting scores* for a landclass disagree with the original expert opinion score (i.e. neither FWHM overlap the expert opinion score uncertainty), the calibrated nesting scores are typically higher than the experts' nesting scores, and this is especially noticeable for crops (Fig. 5). In contrast, where both calibrated *floral scores* for a landclass disagree with the expert opinion scores, the

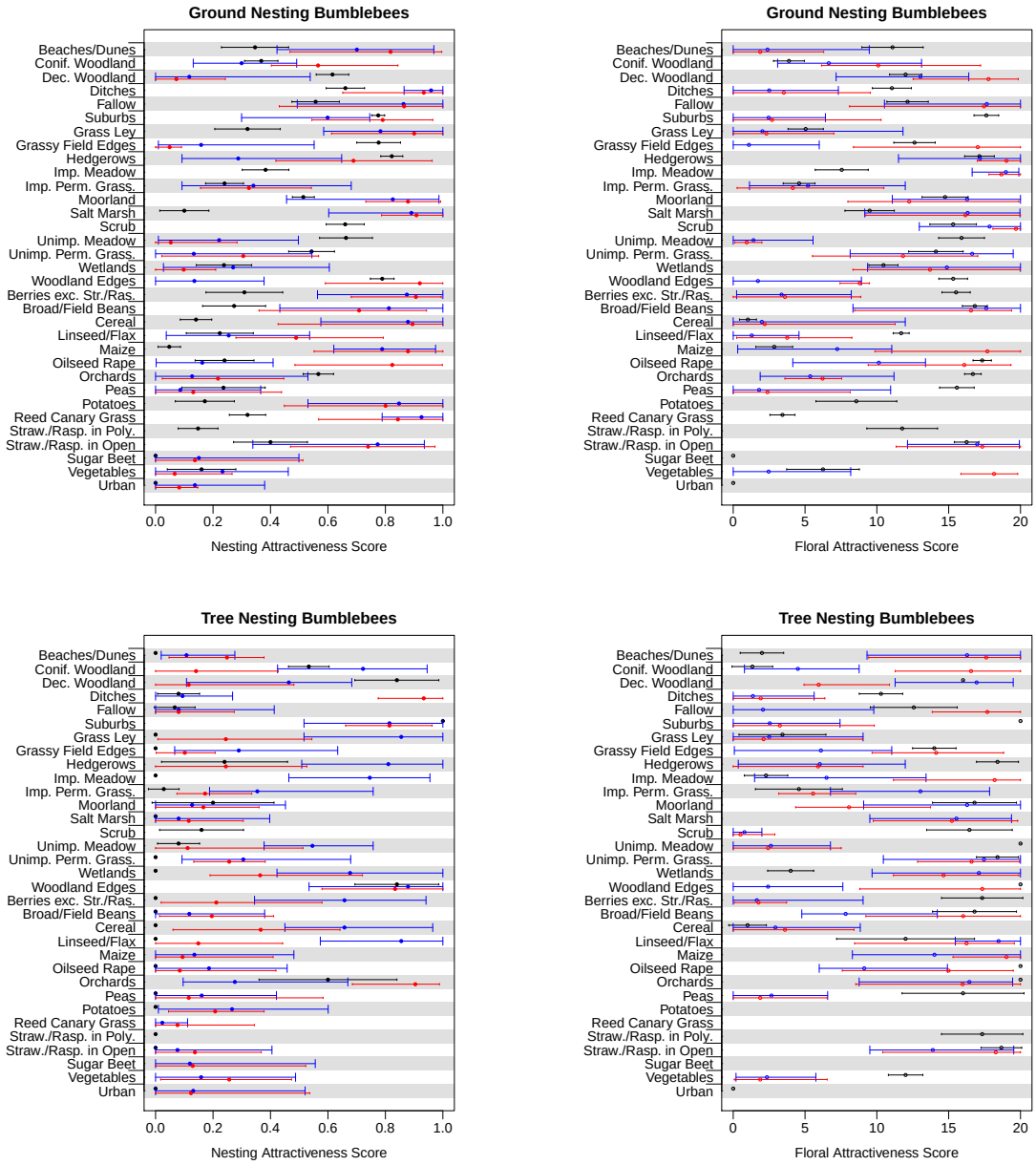

Figure 5: Comparison of expert opinion attractiveness scores (black) with calibrated attractiveness scores obtained by Method 1 (blue) and Method 2 (red) for Ground Nesting Bumblebees and Tree Nesting Bumblebees. Error bars show standard error on expert opinion scores (or zero when only one expert contributed a score or all experts volunteered the same score) and density distribution FWHM for calibrated scores. The absence of a black point for a parameter indicates no experts contributed a score. The absence of a blue (or red) point indicates Method 1 (or Method 2) could not calibrate this parameter. See Table S2 for abbreviations.

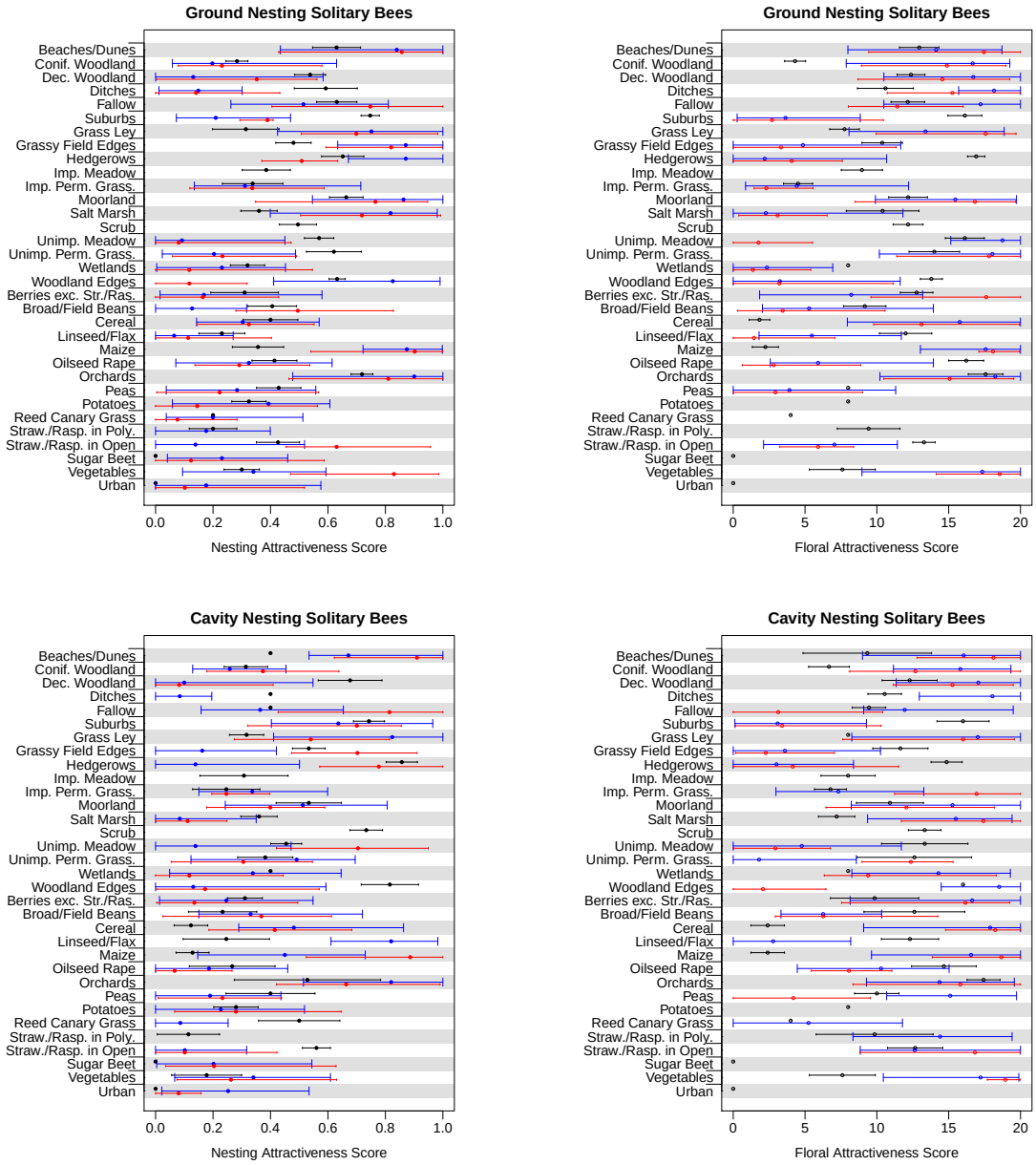

Figure 6: Same as Fig.5 but comparing expert opinion attractiveness scores (black) with calibrated attractiveness scores obtained by Method 1 (blue) and Method 2 (red) for Ground Nesting Solitary Bees and Cavity Nesting Solitary Bees.

calibrated floral scores for bumblebees and GNSB are typically lower than the original expert scores. The solitary bee nesting scores show the greatest level of agreement between the expert opinion scores and calibrated scores (Fig. 6), with 52% of landclasses showing overlapping uncertainties for all three scores compared to 33% and 27% for GNBB and TNBB, respectively.

For individual attractiveness parameters, there can also be large differences between the scores obtained by the two calibration methods. All guilds have instances where the Method 1 calibrated score agrees with the original expert opinion score, while the Method 2 score disagrees, and vice versa. Even though the two calibrated versions of the model produce similar β and R^2 values when compared to the data, Fig. 5 and 6 show that they achieve this sometimes with very different input attractiveness parameters.

Examining the scores for a particular landclass across all guilds reveals some notable trends. The calibrated floral scores for *suburbs* are far lower than the expert floral scores across all guilds (Fig. 5 and 6). The calibrated floral scores for *cereal* and *maize* are significantly higher than the expert floral scores for solitary bees. Finally, the calibrated nesting scores for maize are significantly higher than the expert scores for both ground nesting guilds.

3.4 Calibrated vs Uncalibrated Model Predictions for Great Britain

Fig. 7, and the corresponding Fig. S4, S5 and S6 in the Supplementary Material, show the predicted spring visitation rate across Great Britain for GNBB, TNBB, GNSB and CNSB, respectively, using the nesting and floral attractiveness scores obtained via calibration Method 1. The most extensive regions of predicted high visitation for bumblebees occur in northern Scotland due to large continuous tracts of moorland and wetland (i.e. upland bog in LCM2015) in these areas, which have high floral and nesting calibrated scores for these guilds (Fig. 5). Both bumblebee guilds show lower visitation rates in lowland arable areas of eastern England, due to the predominance of (low calibrated floral score) cereals in these areas (Fig. 7 and S4). However, for GNBB, this low visitation cereal matrix is interspersed with highly visited hedges, fallow and mass flowering crop fields, while TNBB show visitation rate hotspots in East Anglia, where highly scored nesting habitats (deciduous woodland and suburbs) are embedded in highly scored foraging habitats (permanent grassland). The solitary bees show an opposite geographical trend, with higher visitation rates in lowland arable areas (driven by high nesting and floral calibrated scores for maize and cereals; Fig. 6) and lower visitation rates in upland areas of permanent grassland (Fig. S5 and S6).

For all guilds, the spatially resolved uncertainty on the spring visitation rate predictions (top right panels; Fig. 7, S4, S5 and S6) is highest in landclasses where the calibrated attractiveness scores have large FWHM and in landclasses with high floral scores but low nesting scores. Such areas have very little nesting within them, so their visitation rate is dominated by bees nesting in surrounding landclasses. Varying their floral attractiveness therefore has a large effect on how many bees travel in to forage in these areas.

Across all guilds, the calibrated model typically predicts higher visitation rates than the expert model in arable areas of southern, central and eastern England and lower visitation rates than the expert model in upland areas of permanent grassland and suburban areas (bottom left panels; Fig. 7, S4, S5 and S6). TNBB show the most extreme differences between the two models, producing in some upland wetland areas a factor of 10^7 difference in

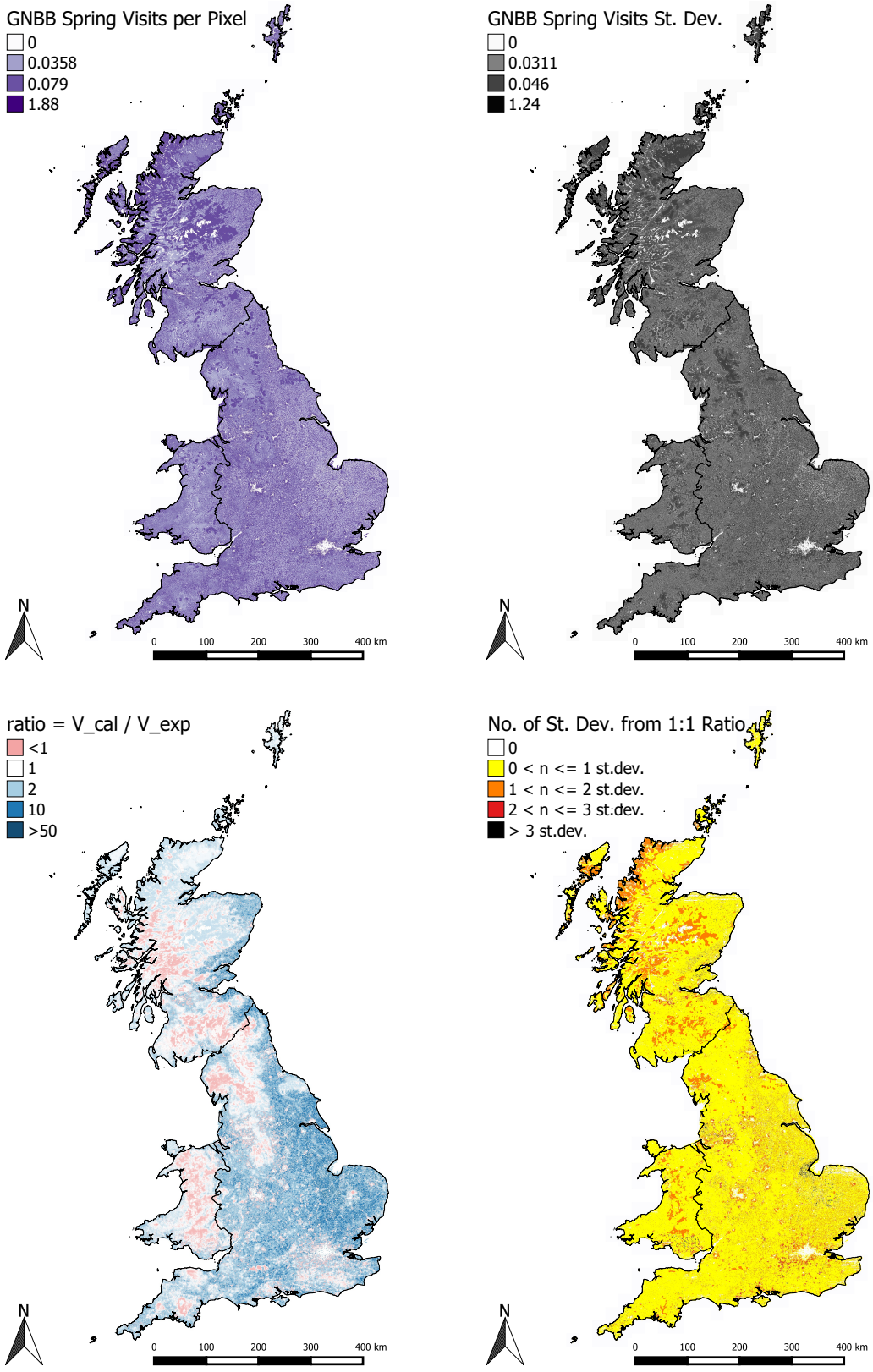

Figure 7: a) Model predictions for Ground Nesting Bumblebee spring visitation rate across Great Britain using attractiveness scores obtained via calibration Method 1. b). Standard deviation of model predictions shown in a., c). Ratio between a. and model predictions using original expert opinion attractiveness scores, d) Number of standard deviations of ratio away from 1:1. Fig. S4, S5 and S6 in the Supplementary Material show corresponding maps for the other pollinator guilds.

predicted visitation rate, due to the calibrated model obtaining non-zero nesting scores for many landclasses for which the experts assigned zero attractiveness (Fig. 5). Once the uncertainties on the attractiveness scores are taken into account, there are generally no significant differences between the two model predictions for solitary bees, with the largest discrepancies (between 1–2 sd away from 1:1 ratio) occurring in suburban areas and at the interface between suburban areas and woodlands (bottom right panels; Fig. S5 and S6). In contrast, both bumblebee guilds show significant differences between the two model predictions in arable areas (> 3sd away from 1:1 ratio; Fig. 7 and S4). See the Supplementary Material for a full discussion of the national-level model predictions for each guild.

4 Discussion

We have compared the most advanced spatially-explicit process-based pollinator abundance model currently available to bee abundance data collected at 239 sites across Great Britain. Our initial model version, parameterised using expert opinion nesting and floral attractiveness scores, showed significant (but non-linear) model-data agreement for all four guilds. We then tested two different methods to calibrate the nesting and floral attractiveness scores for each guild and improve the model-data agreement — 1. a free purely data-driven calibration and 2. an expert-informed prioritised calibration. Method 2 calibrated parameters at a faster rate (initially) for three of the four guilds, but not for GNSB.

Although our calibrated models both showed improvements in model-data agreement, there were significant differences between the calibrated attractiveness scores obtained by the two methods, reflecting the fact that, in complex interacting process-based models, the order in which parameters are calibrated matters. Another factor may be 'over-adjustment' of parameters by the prioritised calibration method to compensate for the fact that the process couldn't always simultaneously adjust other parameters which, if allowed to vary, might have enabled a better combined fit to the data. It may also simply be a consequence of our low model-data sensitivity to small area landclasses (Yapo et al., 1996).

Both calibration processes selected attractiveness scores that improved the fit to our observed abundance data. However, closer examination of the calibrated scores reveals instances where the calibration process identified ecologically unrealistic values, e.g. high floral attractiveness scores for solitary bees for cereals, which are wind pollinated and do not provide significant nectar resources (except potentially in organic systems with higher in-crop wild floral cover; Holzschuh et al., 2007). There are many reasons why our calibration processes might find erroneous/unrealistic attractiveness values:

- **Detectability biases in survey data.** Erroneous calibrated scores may arise if species detectability varies systematically with landcover, e.g. reduced sight lines/fewer individuals in flight. Our guild-level approach may further exacerbate this if guild species composition systematically alters across landcover such that more readily detectable species are replaced with less detectable species in some habitats, so causing an apparent reduction in measured guild abundance unrelated to actual guild abundance. Solitary bees are typically under-recorded on transects due to their smaller size (O'Connor et al., 2019), and their short flight periods also reduces data availability for these guilds.

- **Use of survey totals.** We necessarily compared mean model visitation rate within the survey area with summed abundance along all surveyed transects. This may produce erroneous calibrated scores in heterogeneous survey areas containing multiple landclasses.
- **Timing of crop surveys.** All crop surveys were conducted during the (relatively short) peak flowering period of the crop when temporarily high foraging abundances occur within the crop relative to the wider countryside. However, the model predicts total visitation rate per season. Two adjacent parcels containing equally attractive resources for short durations will receive the same predicted seasonal visitation rate, i.e. half the bees forage in each. If, in reality, these two parcels flower sequentially within one season, such that all the bees forage in one and then in the other, the model cannot capture this unless we subdivide the season and increase the temporal resolution at which we run the model. This temporal limitation may be driving the unusually high calibrated nesting/floral scores obtained for some crops.
- **Geographical distribution of survey sites.** Biases in geographical coverage can produce spurious calibration results if these correlate with systematic changes in landcover or data collection conditions. Despite wide geographical coverage, there were more lowland intensive arable sites than upland sites and often better survey conditions at lowland sites. Wide variation in survey weather condition recording/lack of recording for some sites prevented us controlling for this.
- **Geographical differences in total abundance unrelated to floral/nesting attractiveness.** Climatic gradients and current range limitations (relevant to TNBB *Bombus hypnorum*) can also cause variation in pollinator population size and produce spurious calibration results where these gradients correlate with systematic geographical changes in landcover.
- **Limitations of mapping data: miss-classifications.** Inaccuracies in mapping data can lead to spurious calibration results if mapping data indicate a landclass is present when in reality it is not.
- **Limitations of mapping data: omissions.** Lack of fine-scale feature mapping (Potts et al., 2016) prevents many important pollinator habitats from being included in our input model landscapes. We also only mapped obviously pollinator relevant agri-environment features and used a simplistic approach of placing boundary features around the entire perimeter of the containing land parcel, due to lack of information on feature placement.
- **Limitations of mapping data: No accounting for within-habitat heterogeneity.** Large-scale systematic differences in habitat quality between regions (e.g. due to management) could influence the calibrated attractiveness scores, while small-scale within-habitat heterogeneity will influence measured abundances in the field but won't be present in the mapping data, which is predominantly derived from the 25x25m resolution LCM2015 dataset.
- **Dataset asynchrony and dynamic landscapes.** Crop rotation means that our study landscapes are likely to contain roughly the right proportions of crops but not necessarily in exactly the right places due to asynchrony of our mapping and survey data. Although we forced the surveyed crop fields to contain the correct crop, erroneous attractiveness scores may be obtained for crops that are adjacent in our mapped landscapes but were not adjacent in reality at the time of the survey. Lack of crop rotation information

also means we cannot account for the legacy of past flowering crop distributions on current year pollinator population size/distribution.

- **Non-stationary populations and flight seasons.** We compared the observed data to the predicted spring visitation rates using a survey date co-variable. For bumblebees, this reflects the fact that numbers increase as spring-foraging queens produce summer-foraging workers. The model only permits solitary bees to fly in one season (with no allowance for primitive eusocial or multi-voltine behaviour) and so in order to compare spring visitation rates, we simulated only spring-flying solitaries. By comparing spring solitary bee and bumblebee numbers to survey data collected throughout spring–summer with a date co-variable, we are assuming that spring and summer abundance obey the same correlation over time in different landscapes, which is unlikely to be true if some landscapes contain a high proportion of landclasses with very temporally restricted floral cover scores (Persson and Smith, 2013). An improvement would therefore be to explicitly model both spring- and summer–flying solitary bees and to match surveys to the appropriate seasonal visitation rate. However, this adds an extra layer of complexity to an already complex process and can produce very different results depending on where the (arbitrary and latitude-dependent) cut-off between spring and summer is placed.
- **Choice of parameters to calibrate.** We did not calibrate the floral cover scores, leaving these fixed at their expert opinion estimates to enable decoupling of the guilds. However, experts can struggle to accurately assess floral cover (Baey et al., 2017) and quantitative sampling (e.g. Baude et al., 2016; Hicks et al., 2016) can provide more accurate estimates. Under/over-estimated floral cover scores could cause higher/lower floral attractiveness scores, respectively.
- **Parameter degeneracy.** Crop sites consisted of observational data collected within a single landclass, however, without multiple simultaneous observations in adjacent landclasses with different nesting/floral properties, the calibration process will struggle to disentangle which (i.e. nesting/floral/both) attractiveness scores for the landclasses should be altered to match the data. Measurements in multiple nearby landclasses are needed to capture the movement of bees from good nesting to good foraging areas and so break this degeneracy. This is another certain cause of unrealistic calibrated scores for agricultural landclasses in particular.
- **Structural limitations of model.** The model does not account for density dependent competition for floral resources, land-use factors such as pesticide risk, or flexibility in foraging range. Changes in guild species composition with habitat may cause a change in the typical foraging range for that guild, which may impact calibrated attractiveness scores.

The fact that our expert-informed prioritised calibration process also produced some unrealistic scores raises the question of whether this expert-informed calibration was informed enough. Perhaps we should have gone further and restricted the parameter space searched, e.g. by using the expert opinion scores as Bayesian priors (e.g. Choy et al., 2009). However, the suburban floral attractiveness scores highlight why we might be cautious about taking such a strongly expert-influenced a priori approach; this would potentially have prevented the calibration from even exploring the preferred range identified by both tested calibration methods for all guilds.

There are good reasons why expert opinion scores may be inaccurate or not yield the most appropriate values within our modelling scenario:

- **Expert elicitation method.** Experts scored landclasses independently. A more sophisticated elicitation method, such as the Delphi process (O'Hagan, 2019), may have provided more reliable final scores with lower variance, by allowing the experts to collectively review all opinions and iteratively refine and discuss their scores. In addition, we calculated the certainty-weighted mean score and variance across all experts and used these to parameterise a beta distribution uncertainty profile for each score. Other studies (e.g. Koh et al., 2016) assign beta uncertainty distributions to the individual expert scores and average these, which may yield a broader mean uncertainty distribution and a slightly different weighted mean.
- **Semantic uncertainties.** Each landclass that occurred in the mapping data had to be matched to one of the 35 expert opinion landclasses, generating semantic uncertainties. For example, experts scored 'garden' attractiveness and this was applied to all 'suburbs' in LCM2015, where gardens are diluted by less attractive landclasses e.g. buildings/roads. This could explain the calibrated/expert discrepancy for suburbs. Semantic uncertainties also arise where different experts assign different scores to the same landclass due to different interpretations of a landcover term, e.g. based on field experience in different geographical regions.
- **Knowledge gaps.** There was a trend for the calibrated nesting scores to be higher than the experts predicted. It is difficult to find nests in the field and so plausible that experts in general may be less reliable at assessing nesting quality.

Clearly, we can improve on expert opinion estimates by including data-driven calibration, which relates observational data more directly to the modelling environment (e.g. Groff et al., 2016). However, ecological survey data cannot be treated as 'true' due to its own inherent observational biases. Ideally, expert opinion data would be entirely supplanted with field data on nesting and floral attractiveness, but these require specialised efforts to obtain, are hard to determine (e.g. Osborne et al., 2008; Bahlai and Landis, 2016; Baey et al., 2017) and can vary strongly between species even within guilds (Falk, 2015). The collection of large scale systematic pollinator monitoring data (as proposed by Carvell et al., 2016) could help our data-driven calibration to derive more realistic, consistent estimates with lower temporal/regional biases, but no such data are currently available for the UK. This means some expert moderation is essential to identify unrealistic parameter values, which may reflect the limitations and biases of our current datasets and the insensitivities of our model more than the preferences of the species we are modelling.

Our results show a totally expert opinion parameterisation and a purely data-driven calibration both have limitations in their ability to yield accurate parameter estimates. Our maps of pollinator visitation illustrate how differences in the parameter values obtained by these two approaches can produce enormous differences in model outputs (e.g. factor of 10^7 increase in TNBB visitation in some locations) when used to predict abundance on a landscape scale. This emphasises the need to reconcile these two approaches and obtain the most reliable/realistic estimate for each parameter and, crucially, the approach which yields the most reliable estimate may be different for each parameter. The fact that our expert-informed prioritised calibration also yielded unrealistic parameter values suggests that a more integrated, iterative approach may be better.

We propose a solution is to integrate data-driven calibration results within a Delphi-like process, so adding a data-driven 'expert' to the human members of the panel. Expert opinion is not imposed a priori, but an initial independent data-driven calibration is conducted for comparison with expert opinion. Each calibrated parameter can then be discussed, examining reasons why the data-driven calibration may be preferred over the expert estimate and vice versa. Unrealistic parameter values can be identified, appropriate limits (priors) set if over-adjustment is suspected and the calibration process repeated. Model predictions generated using the final hybrid parameter values can then be compared to the original survey data, to ensure significant agreement is still maintained.

5 Conclusions

Reliably modelling pollinator abundance is essential if we are to identify areas of pollination service deficit and effectively target resources to support pollinator populations. The central place foraging behaviour of many pollinators favours a process-based model in order to accurately reflect how the distribution of nesting/floral resources affects landscape-level pollinator abundance. We selected the most advanced process-based pollinator abundance model available and calibrated it against observational data collected across Great Britain, to assess its suitability for generating spatially-explicit estimates of national pollinator abundance.

In its initial expert-parameterised version, the model showed significant agreement with the survey data, which further improved with calibration for three out of four modelled pollinator guilds. This demonstrates the model's potential to reliably map pollination service/natural capital, identify target areas for interventions and form the basis of novel tools to inform land-use decision-making. Our aim was to identify the parameter set that produced the best fit to the survey data *and* could be used with confidence to predict the consequences of land-use change on UK pollinator populations. Although the calibrated parameterisations satisfy the former, their inclusion of unrealistic parameter values means they fail at the latter; adopting the calibrated parameters for the sake of a small increase in R^2 would (far more seriously) cause the model to predict that increasing cereal cover is beneficial for many pollinators, which is generally not the case. This demonstrates that our concept of model accuracy must include both accurate prediction within the calibration/validation environment *and* ecological realism of underlying parameters (given our wider knowledge base) to enable meaningful model application outside it.

Our work highlights the universal challenges faced when calibrating any spatially-explicit, process-based ecological model. The desire to realistically represent complex ecological processes in finely mapped landscapes necessarily generates models with large numbers of parameters. Computational limitations and model insensitivities may preclude calibration of all parameters making some use of expert estimates a necessity. This, combined with survey and geographical data biases, may lead purely data-driven calibration to easily identify spurious parameter values. We suggest that treating expert elicitation and data-driven calibration as complementary parts of one single iterative process, which integrates model calibration and credibility assessment, may provide the best opportunity to obtain realistic parameter estimates for process-based models, in ecological systems with expert knowledge gaps and patchy/biased ecological data.

6 Acknowledgements

We thank our ten expert elicitation participants for their contribution and the rural payment agencies of England, Scotland and Wales for allowing access to their datasets. We also thank S.Roberts for very helpful discussions on nesting/floral attractiveness scores and R.Sibly for helpful discussions on Approximate Bayesian Computation. This Research is supported by the grant "Modelling Landscapes for Resilient Pollination Services" (BB/R00580X/1), funded by the Global Food Security 'Food System Resilience' Programme, which is supported by BBSRC, NERC, ESRC and the Scottish Government.

7 Author Contributions

EG carried out the research (developed and implemented calibration process, enhanced/adapted ScaLE-poll model, analysed observational datasets, generated maps, conducted analyses) and wrote the manuscript. TB developed and conducted the expert elicitation questionnaire, obtained access to spatial datasets, and assimilated observational datasets. YC provided ScaLE-poll pollinator model and advice on its use. EG, TB, TO, YC and HS developed the modelling approach and interpretation of results, while all other authors contributed observational datasets for use in model calibration. All authors provided comments on the manuscript which were incorporated into the final version.

8 Data Availability

Survey datasets available from original study authors on request (see Table S1).

References

Charlotte Baey, Ullrika Sahlin, Yann Clough, and Henrik G Smith. A model to account for data dependency when estimating floral cover in different land use types over a season. *Environmental and ecological statistics*, 24(4):505–527, 2017.

Christie A Bahlai and Douglas A Landis. Predicting plant attractiveness to pollinators with passive crowdsourcing. *Royal Society open science*, 3(6):150677, 2016.

Nicholas J Balfour, Jeff Ollerton, Maria Clara Castellanos, and Francis LW Ratnieks. British phenological records indicate high diversity and extinction rates among late-summer-flying pollinators. *Biological conservation*, 222: 278–283, 2018.

Mathilde Baude, William E Kunin, Nigel D Boatman, Simon Conyers, Nancy Davies, Mark AK Gillespie, ..., and Jane Memmott. Historical nectar assessment reveals the fall and rise of floral resources in britain. *Nature*, 530 (7588):85–88, 2016.

Matthias A Becher, Volker Grimm, Pernille Thorbek, Juliane Horn, Peter J Kennedy, and Juliet L Osborne.

Beehave: a systems model of honeybee colony dynamics and foraging to explore multifactorial causes of colony failure. *Journal of Applied Ecology*, 51(2):470–482, 2014.

Matthias A Becher, Grace Twiston-Davies, Tim D Penny, Dave Goulson, Ellen L Rotheray, and Juliet L Osborne. Bumble-beehave: A systems model for exploring multifactorial causes of bumblebee decline at individual, colony, population and community level. *Journal of applied ecology*, 55(6):2790–2801, 2018.

T.D. Breeze, A.P. Bailey, K.G. Balcombe, T. Brereton, R Comont, M. Edwards, ..., and C. Carvell. Pollinator monitoring more than pays for itself. (in review), 2019.

Tom D Breeze, Nicola Gallai, Lucas A Garibaldi, and Xui S Li. Economic measures of pollination services: shortcomings and future directions. *Trends in Ecology & Evolution*, 31(12):927–939, 2016.

Luísa Gigante Carvalheiro, William E Kunin, Petr Keil, Jesus Aguirre-Gutiérrez, Willem Nicolaas Ellis, Richard Fox, Quentin Groom, et al. Species richness declines and biotic homogenisation have slowed down for nw-european pollinators and plants. *Ecology letters*, 16(7):870–878, 2013.

Claire Carvell, Nick Isaac, Mark Jitlal, Jodey Peyton, Gary Powney, David Roy, Adam Vanbergen, et al. Design and testing of a national pollinator and pollination monitoring framework. 2016.

Rebecca Chaplin-Kramer, Richard P Sharp, Charlotte Weil, Elena M Bennett, Unai Pascual, Katie K Arkema, Kate A Brauman, et al. Global modeling of nature's contributions to people. *Science*, 366(6462):255–258, 2019.

Samantha Low Choy, Rebecca O'Leary, and Kerrie Mengersen. Elicitation by design in ecology: using expert opinion to inform priors for bayesian statistical models. *Ecology*, 90(1):265–277, 2009.

Steven J Falk. *Field guide to the bees of Great Britain and Ireland*. British Wildlife Publishing, 2015.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

Lucas A Garibaldi, Marcelo A Aizen, Alexandra M Klein, Saul A Cunningham, and Lawrence D Harder. Global growth and stability of agricultural yield decrease with pollinator dependence. *Proceedings of the National Academy of Sciences*, 108(14):5909–5914, 2011.

Michael PD Garratt, Tom D Breeze, Nigel Jenner, Chiara Polce, Jacobus C Biesmeijer, and Simon G Potts. Avoiding a bad apple: Insect pollination enhances fruit quality and economic value. *Agriculture, ecosystems & environment*, 184:34–40, 2014.

Shannon C Groff, Cynthia S Loftin, Frank Drummond, Sara Bushmann, and Brian McGill. Parameterization of the invest crop pollination model to spatially predict abundance of wild blueberry (vaccinium angustifolium aiton) native bee pollinators in maine, usa. *Environmental modelling & software*, 79:1–9, 2016.

Johanna Häussler, Ullrika Sahlin, Charlotte Baey, Henrik G Smith, and Yann Clough. Pollinator population size and pollination ecosystem service responses to enhancing floral and nesting resources. *Ecology and evolution*, 7(6):1898–1908, 2017.

606 Damien M Hicks, Pierre Ouvrard, Katherine CR Baldock, Mathilde Baude, Mark A Goddard, William E Kunin,
607 Nadine Mitschunas, et al. Food for pollinators: quantifying the nectar and pollen resources of urban flower
608 meadows. *PloS one*, 11(6), 2016.

609 Andrea Holzschuh, Ingolf Steffan-Dewenter, David Kleijn, and Teja Tscharntke. Diversity of flower-visiting bees
610 in cereal fields: effects of farming system, landscape composition and regional context. *Journal of Applied*
611 *Ecology*, 44(1):41–49, 2007.

612 Ifan Hughes and Thomas Hase. *Measurements and their uncertainties: a practical guide to modern error analysis*.
613 Oxford University Press, 2010.

614 Christina M Kennedy, Eric Lonsdorf, Maile C Neel, Neal M Williams, Taylor H Ricketts, Rachael Winfree,
615 Riccardo Bommarco, et al. A global quantitative synthesis of local and landscape effects on wild bee pollinators
616 in agroecosystems. *Ecology letters*, 16(5):584–599, 2013.

617 Jeremy T Kerr, Alana Pindar, Paul Galpern, Laurence Packer, Simon G Potts, Stuart M Roberts, Pierre Rasmont,
618 et al. Climate change impacts on bumblebees converge across continents. *Science*, 349(6244):177–180, 2015.

619 Insu Koh, Eric V Lonsdorf, Neal M Williams, Claire Brittain, Rufus Isaacs, Jason Gibbs, and Taylor H Ricketts.
620 Modeling the status, trends, and impacts of wild bee abundance in the united states. *Proceedings of the National*
621 *Academy of Sciences*, 113(1):140–145, 2016.

622 Eric Lonsdorf, Claire Kremen, Taylor Ricketts, Rachael Winfree, Neal Williams, and Sarah Greenleaf. Modelling
623 pollination services across agricultural landscapes. *Annals of botany*, 103(9):1589–1600, 2009.

624 Charlie C Nicholson, Taylor H Ricketts, Insu Koh, Henrik G Smith, Eric V Lonsdorf, and Ola Olsson. Flowering
625 resources distract pollinators from crops: Model predictions from landscape simulations. *Journal of Applied*
626 *Ecology*, 56(3):618–628, 2019.

627 Sandra Nogué, Peter R Long, Amy E Eycott, Lea de Nascimento, José María Fernández-Palacios, Gillian Petrokof-
628 sky, Vigdis Vandvik, and Kathy J Willis. Pollination service delivery for european crops: Challenges and
629 opportunities. *Ecological Economics*, 128:1–7, 2016.

630 Rory S O'Connor, William E Kunin, Michael PD Garratt, Simon G Potts, Helen E Roy, Christopher Andrews,
631 Catherine M Jones, et al. Monitoring insect pollinators and flower visitation: the effectiveness and feasibility
632 of different survey methods. *Methods in Ecology and Evolution*, 2019.

633 Anthony O'Hagan. Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(sup1):
634 69–81, 2019.

635 Ola Olsson, Arvid Bolin, Henrik G Smith, and Eric V Lonsdorf. Modeling pollinating bee visitation rates in
636 heterogeneous landscapes from foraging theory. *Ecological Modelling*, 316:133–143, 2015.

637 Juliet L Osborne, Andrew P Martin, Chris R Shortall, Alan D Todd, Dave Goulson, Mairi E Knight, ..., and
638 Roy A Sanderson. Quantifying and comparing bumblebee nest densities in gardens and countryside habitats.
639 *Journal of applied ecology*, 45(3):784–792, 2008.

Anna S Persson and Henrik G Smith. Seasonal persistence of bumblebee populations is affected by landscape context. *Agriculture, ecosystems & environment*, 165:201–209, 2013.

Simon G Potts, Hien T Ngo, Jacobus C Biesmeijer, Thomas D Breeze, Lynn V Dicks, Lucas A Garibaldi, ..., and Adam Vanbergen. The assessment report of the intergovernmental science-policy platform on biodiversity and ecosystem services on pollinators, pollination and food production. 2016.

Gary D Powney, Claire Carvell, Mike Edwards, Roger KA Morris, Helen E Roy, Ben A Woodcock, and Nick JB Isaac. Widespread losses of pollinating insects in britain. *Nature communications*, 10(1):1018, 2019.

Jörg A Priess, Mathias Mimler, A-M Klein, S Schwarze, T Tscharntke, and I Steffan-Dewenter. Linking deforestation scenarios to pollination services and economic returns in coffee agroforestry systems. *Ecological Applications*, 17(2):407–417, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

Pierre Rasmont, Markus Franzén, Thomas Lecocq, Alexander Harpke, Stuart PM Roberts, Jacobus Christiaan Biesmeijer, Leopoldo Castro, et al. *Climatic risk and distribution atlas of European bumblebees*, volume 10. Pensoft Publishers, 2015.

Taylor H Ricketts and Eric Lonsdorf. Mapping the margin: comparing marginal values of tropical forest remnants for pollination services. *Ecological Applications*, 23(5):1113–1123, 2013.

Maj Rundlöf, Georg KS Andersson, Riccardo Bommarco, Ingemar Fries, Veronica Hederström, Lina Herbertsson, Ove Jonsson, et al. Seed coating with a neonicotinoid insecticide negatively affects wild bees. *Nature*, 521 (7550):77, 2015.

Paul Scholefield, Daniel Morton, Clare Rowland, Peter Henrys, David Howard, and Lisa Norton. Woody linear features framework, great britain v. 1.0. 2016.

Catharina JE Schulp, Sven Lautenbach, and Peter H Verburg. Quantifying and mapping ecosystem services: Demand and supply of pollination in the european union. *Ecological Indicators*, 36:131–141, 2014.

Matthew R Smith, Gitanjali M Singh, Dariush Mozaffarian, and Samuel S Myers. Effects of decreases of animal pollinators on human nutrition and global health: a modelling analysis. *The Lancet*, 386(10007):1964–1972, 2015.

BA Woodcock, JM Bullock, RF Shore, MS Heard, MG Pereira, J Redhead, L Ridding, et al. Country-specific effects of neonicotinoid pesticides on honey bees and wild bees. *Science*, 356(6345):1393–1395, 2017.

Patrice O Yapo, Hoshin Vijai Gupta, and Soroosh Sorooshian. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. *Journal of Hydrology*, 181(1-4):23–48, 1996.

Chang Zhao, Heather A Sander, and Stephen D Hendrix. Wild bees and urban agriculture: assessing pollinator supply and demand across urban landscapes. *Urban Ecosystems*, 22(3):455–470, 2019.

Table 1: ScaLE-poll model parameters taken from literature data showing values adopted for bumblebees (BB; from Häussler et al. 2017), and solitary bees (SB; see Table S3).

Parameter	**Description**	**Unit**	**BB**	**SB**
n_{max}	Number of nests in a cell of maximum nesting quality	nests/ha	19	20
β	Mean dispersal distance for foraging	m	530	191
$\bar{\beta}$	Mean dispersal distance to new nesting sites	m	1000	100
a_w	Median of the growth rate for workers	-	100	-
b_w	Steepness of the growth rate for workers	-	200	-
a_q	Median of the growth rate for reproductive females	-	15000	42
b_q	Steepness of the growth rate for reproductive females	-	30000	12
w_{max}	Max. number of workers produced by a reproductive female	-	600	-
q_{max}	Maximum number of new reproductive females produced	-	160	2
p_w	Fraction of foraging workers	-	0.5	-

Table 2: Results from fitting equation $\log\left(\frac{N_{obs}+1}{L}\right) = \beta \log V_1 + \gamma \log W + \left(\alpha_{2011}, ..., \alpha_{2016}\right) Y$ to assess model-data agreement for initial model predictions using expert opinion attractiveness scores and model predictions using calibrated attractiveness scores obtained via Methods 1 and 2. Statistically significant coefficients are marked with asterisks (*** = $P < 0.001$). Guild abbreviations GNBB, TNBB, GNSB and CNSB refer to ground nesting bumblebees, tree nesting bumblebees, ground nesting solitary bees and cavity nesting solitary bees, respectively.

		GNBB	**TNBB**	**GNSB**	**CNSB**
Expert Opinion	β	0.23 ± 0.07 ***	0.16 ± 0.02 ***	0.50 ± 0.05 ***	0.44 ± 0.04 ***
	γ	-0.2 ± 0.2	-0.62 ± 0.09 ***	-0.8 ± 0.1 ***	-0.62 ± 0.08 ***
	R^2	0.298	0.448	0.285	0.467
Calibrated Method 1	β	0.91 ± 0.06 ***	0.75 ± 0.05 ***	0.99 ± 0.06 ***	0.89 ± 0.05 ***
(free data-driven)	γ	0.1 ± 0.1	-0.35 ± 0.09 ***	-0.7 ± 0.1 ***	-0.49 ± 0.07 ***
	R^2	0.358	0.460	0.382	0.482
Calibrated Method 2	β	0.83 ± 0.06 ***	0.84 ± 0.03 ***	0.90 ± 0.06 ***	0.83 ± 0.05 ***
(expert-informed	γ	0.1 ± 0.1	-0.09 ± 0.08	-0.8 ± 0.1 ***	-0.58 ± 0.08 ***
prioritised)	R^2	0.342	0.433	0.377	0.486