Received: 11 November 2020 | Accepted: 5 November 2021

DOI: 10.1111/1365-2664.14127

Journal of Applied Ecology | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Predicting spatio-temporal distributions of migratory populations using Gaussian process modelling

Antti Piironen | **Juho Piironen** | **Toni Laaksonen**

University of Turku, Turku, Finland

Correspondence
Antti Piironen
Email: antti.p.piironen@utu.fi

Funding information
Ministry of Agriculture and Forestry

Handling Editor: Ayesha Tulloch

Abstract

1. Knowledge concerning spatio-temporal distributions of populations is a prerequisite for successful conservation and management of migratory animals. Achieving cost-effective monitoring of large-scale movements is often difficult due to lack of effective and inexpensive methods.
2. Taiga bean goose *Anser fabalis fabalis* and tundra bean goose *A. f. rossicus* offer an excellent example of a challenging management situation with harvested migratory populations. The subspecies have different conservation statuses and population trends. However, their distribution overlaps during migration to an unknown extent, which, together with their similar appearance, has created a conservation–management dilemma.
3. Gaussian process (GP) models are widely adopted in the field of statistics and machine learning, but have seldom been applied in ecology so far. We introduce the R package `gplite` for GP modelling and use it in our case study together with birdwatcher observation data to study spatio-temporal differences between bean goose subspecies during migration in Finland in 2011–2019.
4. We demonstrate that GP modelling offers a flexible and effective tool for analysing heterogeneous data collected by citizens. The analysis reveals spatial and temporal distribution differences between the two bean goose subspecies in Finland. Taiga bean goose migrates through the entire country, whereas tundra bean goose occurs only in a small area in south-eastern Finland and migrates later than taiga bean goose.
5. *Synthesis and applications*. Within the studied bean goose populations, harvest can be targeted at abundant tundra bean goose by restricting hunting to south-eastern Finland and to the end of the migration period. In general, our approach combining citizen science data with GP modelling can be applied to study spatio-temporal distributions of various populations and thus help in solving challenging management situations. The introduced R package `gplite` can be applied

Antti Piironen and Juho Piironen contributed equally.

not only to ecological modelling, but to a wide range of analyses in other fields of science.

KEYWORDS
adaptive management, citizen science, distribution modelling, ecological modelling, flyway management, migration ecology, waterfowl ecology, wildlife management

1 | INTRODUCTION

Human population growth and the intensifying use of natural resources pose increasing challenges to the conservation and management of wildlife populations (e.g. Halpern et al., 2008). Consequently, national and international activities have been proposed and initiated to safeguard the sustainable use and preservation of wildlife populations (e.g. Hawkins et al., 1984). Decision-making in conservation and management requires reliable data on population dynamics and ecosystem processes, but relevant information is often scarce, emphasizing the importance of using all available data with suitable statistical tools (Johnson et al., 2018).

Knowledge of distribution over the annual cycle is a prerequisite for the successful conservation and management of migratory animals, as it plays a vital role in habitat safeguarding, population monitoring and targeting management actions. Understanding spatio-temporal dispersion is particularly important in cases where multiple populations with different conservation statuses occur in the same area and are affected by the same human actions. Birds are probably the best-known migratory animals, but despite their movements having been intensively studied since bird ringing began in the late 1800s, the spatio-temporal occurrences of many species and populations are still poorly understood. Traditional methods, such as bird ringing, are usually ineffective and slow (Anderson & Green, 2009), whereas modern tracking technologies suffer from expensiveness, the large size of tracking devices or short life span of small devices (Tomkiewicz et al., 2010).

Citizen science may offer valuable tools for nature conservation and management (McKinley et al., 2017), but the data often suffer from weaknesses caused by spatial and temporal observation biases or the insufficient expertise of observers (Callaghan et al., 2019). Producing scientific knowledge from these large yet heterogeneous datasets often requires applying modern statistical methods in the analyses. Unfortunately, commonly used methods often have many weaknesses with heterogeneous data collected by citizens (Bird et al., 2014). Gaussian processes (GPs) offer a flexible probabilistic approach for modelling such data. The basic theory has been known for decades (e.g. O'Hagan, 1978), and the machine learning community became aware of GPs in the 1990 (e.g. Williams & Rasmussen, 1996), and nowadays they are commonplace in the field (for an excellent introduction, see Rasmussen & Williams, 2006). Previous applications in ecology, however, are relatively sparse. GPs have been used to study optimization in fisheries and predator–prey interactions (Patil, 2007), species distribution modelling (SDM; Vanhatalo et al., 2012; Golding & Purse, 2016; Ingram et al., 2020; Vanhatalo et al., 2020; Wright et al., 2021), modelling individual fish growth (Sigourney et al., 2012) and decision-making in fisheries (Boettiger et al., 2015). The GP models' flexibility and capability to account for uncertainties due to geographically and temporally uneven observation pressure enable wider usage in ecology. So far, their applicability has undoubtedly been limited by the absence of user-friendly tools for the R language, which is the de facto programming language in the field. Many R packages that provide some GP regression functionalities are limited in features, and do not support functionalities necessary for general-purpose modelling (e.g. packages `gptk`, `mlegp` and `GPfit` all implement only Gaussian noise model).

Migratory waterfowl are excellent examples of difficult conservation–management situations, as they are important quarry species but many of their populations have declined in recent decades (Madsen et al., 2015). Simultaneously, other populations, even sympatric ones, are so abundant that they require population control (Fox & Madsen, 2017). Species with different population trends can be affected by the same management actions (e.g. hunting, habitat management). For example, various waterfowl species are often similar in appearance and thus difficult to identify in a hunting situation, which complicates their harvest management. Difficult management situations with two sympatric, look-a-like birds with opposite conservation statuses have been recognized in North America (e.g. Sheaffer et al., 2004), where diverse management challenges have been dealt with by applying an adaptive harvest management framework since the 1990s (Nichols et al., 2007). In Europe, adaptive management approaches for waterfowl have been introduced more recently (e.g. Madsen et al., 2017).

The two Western Palearctic subspecies of bean goose, the taiga bean goose *Anser fabalis fabalis* and the tundra bean goose *Anser fabalis rossicus* provide an excellent example of a within-species conservation–management dilemma. The tundra bean goose population has doubled since the late 1980s and is recently estimated at 600,000–650,000 individuals (Heinicke, 2018). In contrast to that, taiga bean goose numbers have decreased in recent decades, with latest population estimates reaching 70,000–80,000 individuals (Heldbjerg et al., 2019). Both bean goose subspecies are legal quarry in many countries within their range, but due to their different population statuses and trends, their conservation and management needs are clearly different. Unfortunately, they are very similar in appearance and therefore impossible to identify in a hunting situation. This leads to considerable difficulties when aiming to target

the harvest towards the abundant tundra bean goose without hampering taiga bean goose conservation goals. Taiga bean geese breed in the boreal zone of Fennoscandia and north-western Russia, and winter mainly in southern Sweden, northern Germany and Poland. Tundra bean geese breed in the tundra zone and winter in a broad area in western and central Europe (see Figure 1). However, the subspecies can overlap in their migration stop-over areas. The movements of taiga bean geese breeding in Finland are fairly well known (e.g. Nilsson, 2011), but tundra bean geese occurring alongside taiga bean geese on their migration through Finland is poorly understood. Honka et al. (2017) showed, using molecular genetic methods, that bean geese harvested in south-eastern Finland were mainly tundra bean geese, whereas birds from western and northern Finland were mainly taiga bean geese. Nonetheless, this information is coarse due to the small sample size (N = 103). Additionally, Honka et al.'s (2017) study was lacking the temporal component, meaning that the study did not account for yearly variation in subspecies occurrence.

Knowledge concerning spatio-temporal differences of bean goose subspecies occurrences may enable geographical and seasonal hunting regulations and thus prevent overharvesting of taiga bean geese. Despite the look-alike problem that makes bean goose subspecies identification impossible in a hunting situation, taiga and tundra bean geese have certain characteristics that differentiate the appearance of their heads and bills (Heinicke, 2010). These characteristics allow subspecies identification for most individuals in the field with a spotting scope, and thus enable birdwatchers to collect bean goose observations on a subspecies level.

The aim of this paper is to introduce and promote GP modelling as a tool for utilizing citizen science data for studying the spatio-temporal occurrence of migratory populations. In a case study, we apply GP modelling with birdwatcher observation data to predict differences in taiga and tundra bean goose spatio-temporal distributions in Finland during migration. As a result, we provide a general-purpose R package `gplite` (Piironen, 2021b) for future GP analyses along with management recommendations for the bean goose management–conservation issue.

2 | MATERIALS AND METHODS

2.1 | Bird observation data

We received birdwatcher observation data collected during 2011–2019 from BirdLife Finland. Observations were collected via the online bird observation portal Tiira (https://www.tiira.fi/). Species, location, date and number of observed birds are mandatory information to the observation. Additionally, observers can save a variety of information such as age, status etc. to the observation. A bean goose observation can be entered into the system as a taiga or a tundra bean goose, or as a bean goose if the subspecies was not identified. In our analysis, we only used observations where the subspecies was identified and where the bird's status was recorded as local (i.e. not flying), as subspecies identification from a flying bean goose is unreliable. We sorted observations from 1.3. to 31.5. for spring migration and from 1.8. to 30.11. for autumn migration annually. In the end, we had *c*. 19,500 observations that met the above-mentioned criteria (See Table S3 in the Supplementary Information). In each observation, 1–12,500 individuals were observed, with a mean of 148. Any ethical approvals for collection of birdwatcher observations were not required.

Our data contain two main sources of uncertainty, both typical to citizen science data. First, the observation effort is not evenly distributed spatially or temporally. This is not a problem, as it will only increase the uncertainty of the model predictions in regions and times with few observations (note that we are interested only in the subspecies ratio, see below). Second, observations are made by numerous birdwatchers with unknown and variable expertise, possibly generating incorrectly identified birds into the data. Nonetheless, the low percentage (*c*. 40%) of bean goose observations identified to subspecies level among all bean goose observations indicate that birdwatchers are somewhat prudent in difficult identification situations, and the majority report their observations only when they are confident with the identification. Additionally, temporal differences in subspecies composition in the same area indicate that no obvious or severe spatial biases

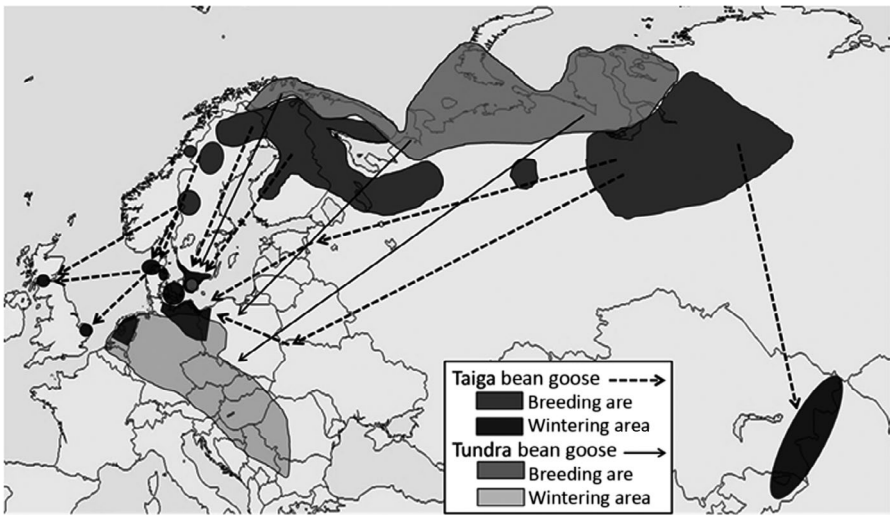

FIGURE 1 Breeding and wintering ranges and approximate migration routes of taiga and tundra bean goose in the Western Palaearctic and western parts of the Eastern Palaearctic according to Marjakangas et al. (2015)

exist in subspecies identification. Therefore, we consider erroneous observations to be randomly distributed in the data (see also Bradter et al. (2018) for a comparison between birdwatcher and systematically collected data).

Instead of modelling the occurrence of either one of the subspecies alone, we modelled the ratio of the two subspecies, as it is advantageous for statistical reasons and for producing better management recommendations. From a statistical viewpoint, modelling the ratio (presence–absence data) is considerably less vulnerable to possible biases originating from the spatially and temporally varying observation effort than modelling the distribution of each subspecies alone (presence-only data). A lack of observations does not introduce bias into estimating the ratio, as it only affects the model's uncertainty concerning the estimate. On the other hand, when modelling the distribution of one subspecies alone, you cannot ignore the bias coming from uneven observation effort, which would complicate the analysis considerably. Regarding management recommendations, our goal is to find a management solution that provides an optimal compromise between taiga bean goose conservation and avoiding unnecessary harvest regulations for tundra bean geese. Thereby, it is vital to recognize times and areas where the proportion of tundra bean geese out of all bean geese is large, as these are the times and areas where harvest is targeted at tundra bean geese while taiga bean geese are spared. Nevertheless, when modelling the ratio, we need to assume that reporting rates between the two subspecies do not differ when they are detected and identified. We are not aware of any reasons why this assumption could not be made.

2.2 | Model

Gaussian processes offer a powerful and flexible way of incorporating prior knowledge into the model while allowing a principled way to handle uncertainties. For a thorough introduction to GPs, we highly recommend the book by Rasmussen and Williams (2006). For a reader new to GPs, we provide some more details in the Supporting Information. Here we shall only give a brief description of the model. For motivation, discussion and comparison to other potential modelling choices, see the Supporting Information.

As discussed in Section 2.1, our data consist of approximately $N = 19{,}500$ observations collected during 2011–2019. The relevant information for our spatio-temporal model from each observation is as follows: coordinates $\mathbf{x} = (x_1, x_2)$, time stamp t (date) and the number of taiga and tundra bean geese observed. We use the symbol $\mathbf{z}$ to denote all the predictor features, $\mathbf{z} = (\mathbf{x}, t) = (x_1, x_2, t)$. For notational convenience, we denote the number of subspecies taiga bean goose in each observation with y, and the total number of birds in the corresponding observation with n. So, for example, $y = 90$ and $n = 100$ mean that 90 taiga and 10 tundra bean geese were observed in that particular event. We note that the data are considered presence–absence in the sense that we assume that both subspecies to be recorded, if either one of them is observed (i.e. an observation of 50 tundra bean geese means that 50 tundra bean geese were observed, but no taiga bean geese). Total absence observations are not assumed to be made.

The data have very obvious overdispersion due to the flocking behaviour of the geese. In fact, only one of the two subspecies was present in *c*. 90% of the observations. To account for this, we assume each observation y_i follows a beta-binomial distribution, which can be written as

$$\begin{aligned} y_i \mid p_i &\sim \text{Binomial}(p_i, n_i), \\ p_i \mid \mu_i &\sim \text{Beta}(a_i, b_i), \\ a_i = \frac{\mu_i}{\phi}&, b_i = \frac{1-\mu_i}{\phi}. \end{aligned} \tag{1}$$

Here parameter $\mu_i \in (0, 1)$ is of the central interest, as it determines the expected value of y_i: $\mathrm{E}(y_i) = n_i \mathrm{E}(p_i) = \mu_i n_i$. What makes this model different from the binomial distribution is the overdispersion parameter ϕ, which increases the variance of y_i compared to the binomial model whenever $\phi > 0$. As $\phi \to 0$, the model approaches $y_i \sim \text{Binomial}(\mu_i, n_i)$.

We model $\mu = \mu(\mathbf{z}) \in (0, 1)$ by introducing a latent function $f = f(\mathbf{z}) \in (-\infty, \infty)$ for which we give a zero mean GP prior, and then transform that through a logistic sigmoid to get μ:

$$\mu(\mathbf{z}) = \frac{1}{1+\exp(-f(\mathbf{z}))}, \quad f(\mathbf{z}) \sim GP\left(0, k(\mathbf{z}, \mathbf{z}')\right). \tag{2}$$

The heart of a GP model is the covariance function (or kernel) $k(\mathbf{z}, \mathbf{z}')$, which specifies the properties of the model. We use the following structure

$$k(\mathbf{z}, \mathbf{z}') \propto k_s(\mathbf{x}, \mathbf{x}') k_t(t, t'). \tag{3}$$

In other words, the covariance factors into spatial and temporal components, which makes it easy to specify both components separately. The multiplicative covariance introduces an interaction between the spatial and temporal variation, meaning that the model allows the latent function to have spatial variation dependent on time. Recall that covariance of the form $k_s(\mathbf{x}, \mathbf{x}') k_t(t, t')$ corresponds to the functional form $f(\mathbf{x}, t) = f_s(\mathbf{x}) f_t(t)$ for the latent function (see Rasmussen & Williams, 2006, Section 4.2.4). An additive covariance structure $k_s(\mathbf{x}, \mathbf{x}') + k_t(t, t')$ (which corresponds to the form $f(\mathbf{x}, t) = f_s(\mathbf{x}) + f_t(t)$) could also be considered so that the latent function would look spatially the same at every time t, but this turned out be a clearly inferior choice in terms of data fit (see Supporting Information for model assessment).

For the spatial component, we use the so-called neural network covariance function (Williams, 1998)

$$k_s(\mathbf{x}, \mathbf{x}') = k_{nn}(\mathbf{x}, \mathbf{x}'). \tag{4}$$

The actual functional form is given in the Supporting Information. The neural network covariance function produces smooth non-stationary functions and has a reasonably good extrapolation ability.

Recall that covariance function $k(\mathbf{x}, \mathbf{x}')$ is said to be stationary if it depends on $\mathbf{x} - \mathbf{x}'$ only, meaning that the latent function f is assumed to vary at the same speed everywhere (see Rasmussen & Williams, 2006, Section 4.2.1). In contrast, the non-stationary neural network kernel allows the latent function to vary more rapidly in the middle and more slowly on the boundaries of the input space $\mathbf{x}$, which matches nicely with our prior beliefs concerning the spatial behaviour of f. The covariance function has two hyperparameters, namely τ_0 and τ, which determine how rapidly f varies, and the wideness of the region where f varies substantially.

As we expect different years to be at least roughly similar, we include a periodic kernel k_{periodic} with period $T = 365$ days in the temporal component. This is achieved using a transformation $\tilde{\mathbf{t}} = \left(\sin\frac{2\pi t}{T}, \cos\frac{2\pi t}{T}\right)$ and then feeding this into some base kernel. As a base kernel, we again use the neural network covariance function, so the periodic component can be written as

$$k_{\text{periodic}}(t, t') = k_{\text{nn}}\left(\tilde{\mathbf{t}}, \tilde{\mathbf{t}}'\right). \tag{5}$$

To allow for some inter-annual variability (i.e. deviation from exact periodicity), we modulate the above kernel by a squared exponential kernel, $k_{\text{se}}(t, t')$ (see Supporting Information for the functional form), whose length–scale hyperparameter ℓ will determine how quasi-periodic the temporal variation is ($\ell \to \infty$ indicating exact periodicity). The magnitude hyperparameter σ_f^2 on the other hand will determine the overall magnitude of variation in the latent function. Thus, the temporal covariance function becomes

$$k_{\text{t}}(t, t') = k_{\text{se}}(t, t')\, k_{\text{periodic}}(t, t'). \tag{6}$$

Combining all the pieces together, we can write the full covariance function as

$$k(\mathbf{z}, \mathbf{z}) = k_{\text{s}}(\mathbf{x}, \mathbf{x}')\, k_{\text{t}}(t, t') \tag{7}$$

$$= k_{\text{nn}}^{(1)}(\mathbf{x}, \mathbf{x}')\, k_{\text{se}}(t, t')\, k_{\text{nn}}^{(2)}\left(\tilde{\mathbf{t}}, \tilde{\mathbf{t}}'\right), \tag{8}$$

The superscripts in the two neural network kernels indicate that there are two separate kernels with similar functional form but separate hyperparameters (and inputs). In total, there are six kernel hyperparameters $\left(\tau_0^{(1)}, \tau^{(1)}, \tau_0^{(2)}, \tau^{(2)}, \ell, \sigma_f\right)$ and one likelihood hyperparameter ϕ in the model. For $\tau_0^{(1)}, \tau_0^{(2)}, \tau^{(1)}$ and $\tau^{(2)}$, we use half-Cauchy priors with unit scale. Other hyperparameters are given log-uniform priors.

For fitting the models, we use the R package `gplite`. The installation instructions and a quick-start tutorial for the package are available at https://github.com/jpiironen/gplite. An example code used for the case study of the present paper is available at https://github.com/jpiironen/anser_fabalis. We fit models separately for spring and autumn, which have approximately 15,700 and 3,800 observations respectively. Both models are identical in design but are fitted separately to the two datasets. The number of observations prohibits the use of a full GP, and we use the fully independent training and test conditional (FITC) approximation with 200 inducing points (Quiñonero-Candela & Rasmussen, 2005; Snelson & Ghahramani, 2006). Due to the non-Gaussian likelihood, approximate inference for the latent values must also be used, and we employ Laplace approximation. Hyperparameters are estimated by optimizing them to their marginal maximum a posteriori values.

3 | RESULTS

Figures 2 and 3 show the model fit and data for the autumn and spring migrations, respectively, on average and across several years. The contours show how the probability for an observed bean goose to be a taiga bean goose (i.e. posterior mean of μ) varies over time at different spatial locations (see caption for more details). As shown in Figure 2, the probability of a bean goose being a taiga bean goose is high throughout Finland at the beginning of migration. Later in autumn, the probability for tundra bean goose increases, especially in south-eastern Finland. However, between-year variation exists in the proportion of tundra bean geese.

Analogous to autumn, our model predicts a high probability for a bean goose to be a taiga bean goose at the beginning of spring migration (Figure 3), whereas the probability of tundra bean goose increases during spring in southern Finland. It is noteworthy that in spring, the main division between subspecies occurrences is in the south–north direction, while being mainly in a south-easterly to north-westerly direction in autumn.

The number of bean geese decreases at the end of both migration periods, which decreases the number of observations. This can be seen as increasing uncertainty in the model predictions (i.e. smaller coverage of shaded grey area in Figures 2 and 3).

4 | DISCUSSION

The aim of our study was to introduce and promote GP modelling as a tool for predicting the spatio-temporal distribution of migratory populations using heterogeneous citizen science data. For these purposes, we introduced the R package `gplite` and demonstrated its use with a case study that analysed spatial and temporal differences in the occurrence of taiga and tundra bean goose in Finland. In the case study, the model predicts significant tundra bean goose occurrence only in south-eastern Finland for both spring and autumn. The width of the area where tundra bean goose occurs varies between years, possibly caused by wind conditions and available food supplies on the fields during migration. Tundra bean goose occurrence is also restricted to a smaller zone during autumn than during spring. These results are compatible with results of molecular genetic study by Honka et al. (2017), who showed that the bean goose hunting bag in eastern Finland contains more tundra than taiga bean geese and vice versa in western Finland. The temporal component was absent in previous work by

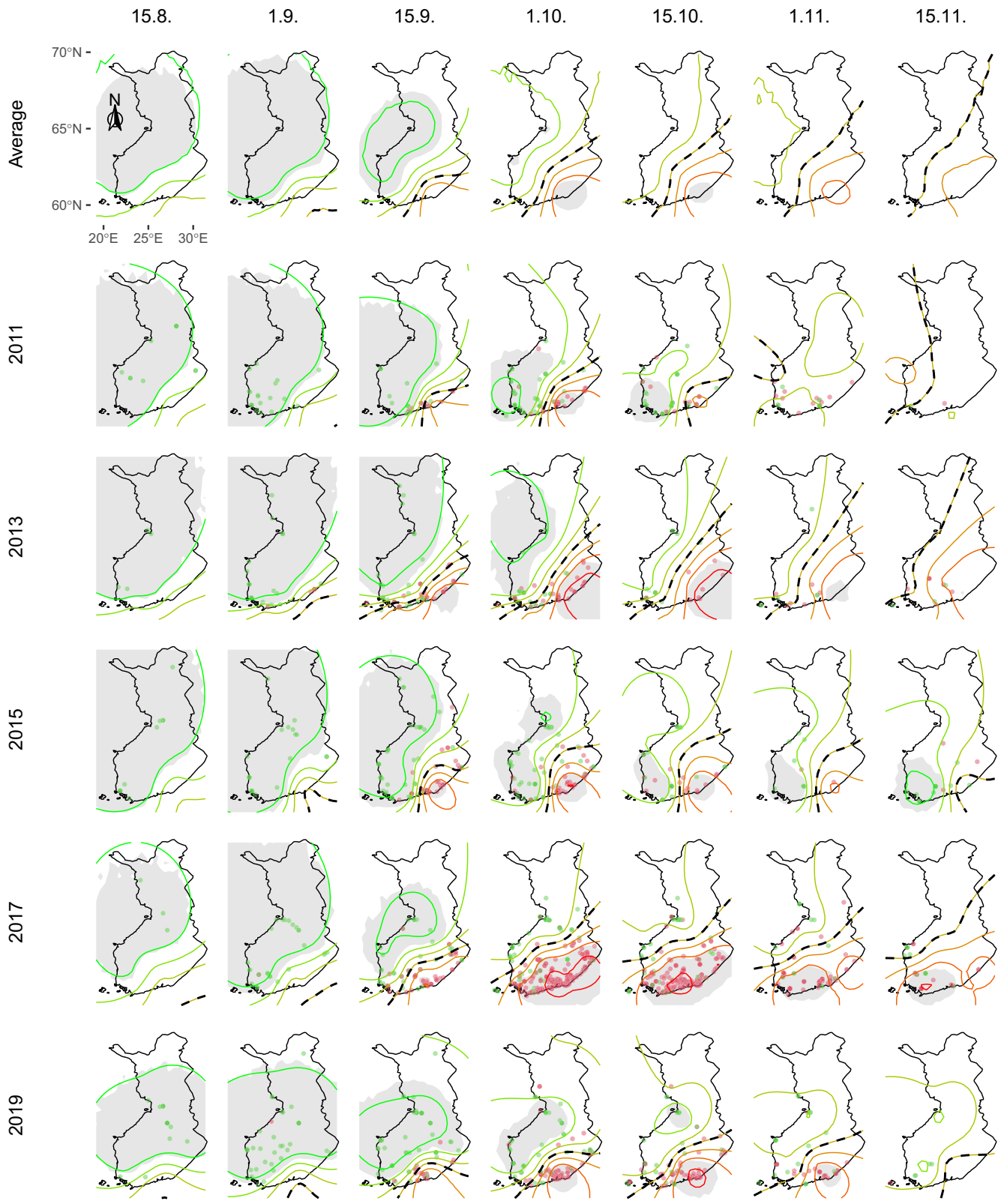

FIGURE 2 Model predictions during the autumn migration at different dates across different years. Due to long time series, only average and every other year is shown in the picture (for full time series, please see Supporting Information). The contours denote the posterior mean for μ (i.e. probability of taiga bean goose) ranging from 0.1 (red) to 0.9 (green) with approximate contour interval 0.114. Dashed black highlights contour $\mu = 0.5$. The same colour denotes the same value for μ throughout the picture. Shaded grey denotes areas where μ is different from 0.5, with posterior probability at least 95%. Dots denote observations within $\pm$ 8 days from the given day; red and green colours mark whether the majority of the observed bean geese were tundra or taiga bean geese respectively

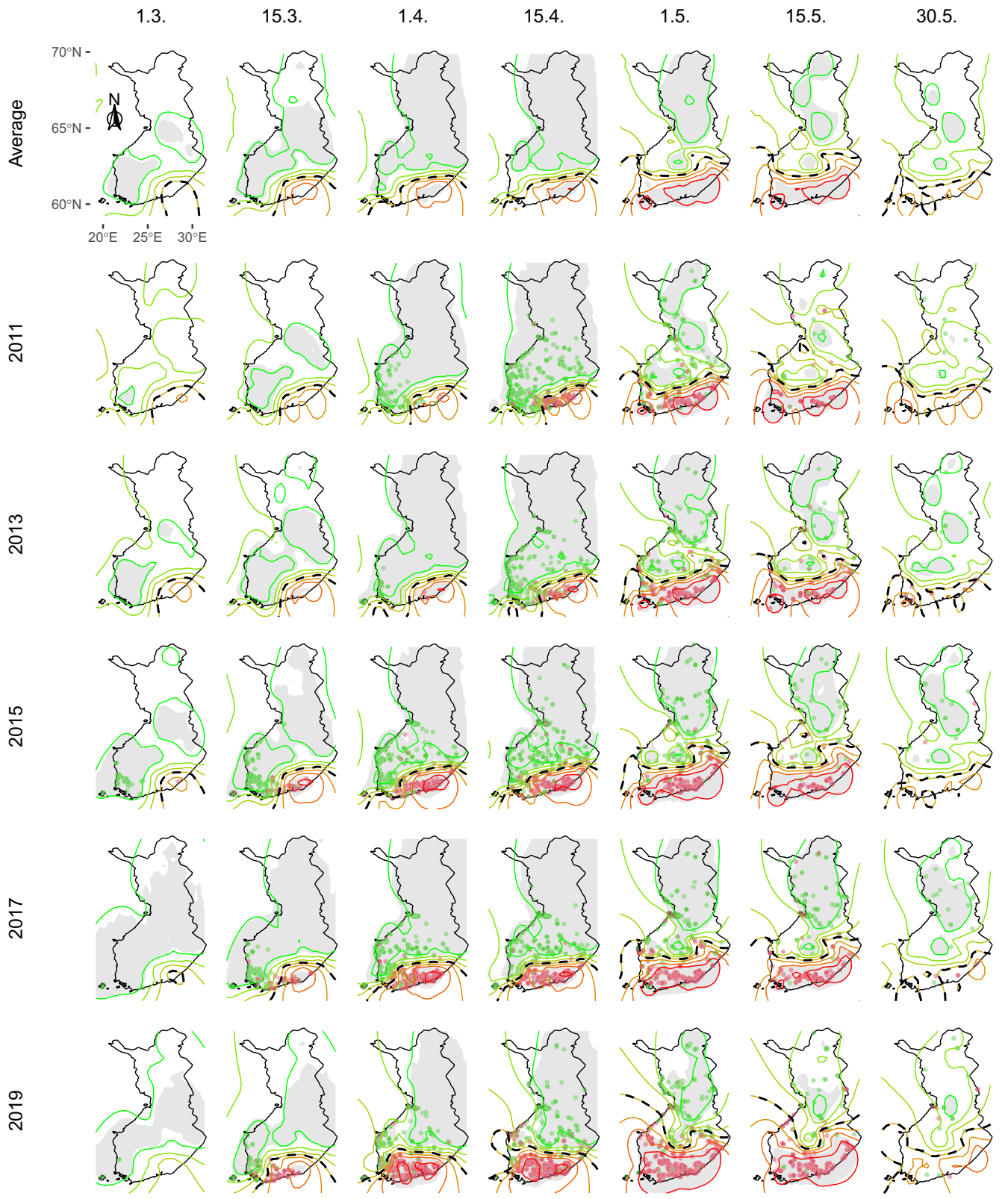

FIGURE 3 Same as in Figure 2, but for the spring migration

Honka et al. (2017). Our approach is the first detailed description of the pattern including both spatial and temporal differences in subspecies occurrence. Our study is also the first one that can be directly applied to the harvest management of these spatially overlapping populations with different population statuses and trends.

4.1 | Differences in bean goose subspecies occurrence

Tundra bean geese migrate later than taiga bean geese, both in spring and autumn. They are almost absent during the beginning of autumn migration in late August and early September, but their proportion of all bean geese in south-eastern Finland increases in the last half of September and remains high in October. Bean goose numbers and consequently the number of bean goose observations decrease in late October, also reducing our model's capability to predict subspecies occurrence (see Figure 2). During spring migration, the taiga bean goose migration begins in the first half of March, when tundra bean geese are nearly absent in Finland. Tundra bean goose numbers begin increasing in the middle of April, and the migration peak occurs approximately at the shift from April into May. Bean goose numbers decrease in the middle of May in Finland (see Figure 3). As discussed earlier in Section 3, the geographical distributions of both subspecies in Finland differ between spring and autumn migration. This is surprising, as it shows that tundra bean goose migration routes and staging areas differ between spring and autumn migration.

4.2 | Advantages of Gaussian process modelling

Previously, a variety of methods have been used to analyse species' spatial and/or temporal distributions with citizen science data. These methods include GLMs (Cheng et al., 2019), occupancy models (Altwegg & Nichols, 2019), maximum entropy models (Phillips et al., 2006) and generalized additive models (GAMs) typically using splines as basis functions (Bird et al., 2014). Tree-based models have also been used, in particular random forests (Prasad et al., 2006) and gradient boosted trees (Elith et al., 2008). Compared to these more conventional approaches, GPs have performed better in terms of predictive accuracy in comparative studies (Golding & Purse, 2016; Ingram et al., 2020; Wright et al., 2021). Additionally, GPs offer a richer (compared to GAMs and GLMs) and more flexible (compared to maximum entropy models) class of models. This includes enhanced ways of incorporating prior knowledge into the model structure and a well-calibrated uncertainty estimation (compared to tree-based models, Hastie et al., 2009, Ch. 9–10).

When comparing GPs to GLMs and GAMs technically, it is well known that GAMs include GLMs as a special case (Hastie & Tibrishanhi, 1990). Analogously, many GAMs, including some based on splines, can be seen as a special case of GPs with a specific covariance function (Rasmussen & Williams, 2006, ch. 6.3 and references therein). Therefore, GPs are inherently a richer class of models, which allow for more flexible model construction through covariance function specification. For low-dimensional data with simple covariance functions such as the squared exponential, the differences between a spline GAM and a GP can be small in interpolation (see, e.g. Riutort-Mayol et al., 2020). However, the ability to add much richer structure to the covariance function (such as quasi-periodicity, non-stationarity, etc.) that affects the model predictions (both in interpolation and extrapolation) is one of the key benefits of GP modelling over spline models. In practical applications, this allows for more complex interactions between features through a more diverse covariance function specification, which can be advantageous in terms of predictive accuracy in various modelling tasks. For practical examples in ecology (in these cases, SDM), see Golding and Purse (2016), Ingram et al. (2020) and Wright et al. (2021).

In a technical comparison to maximum entropy models and tree-based models, GPs differ more fundamentally. Maximum entropy models are designed for SDM under the assumption of presence-only data (Elith et al., 2010), and are therefore inapplicable in studies such as ours. Tree-based models, although as powerful as off-the-shelf models for prediction, suffer from difficulties in incorporating certain types of prior assumptions into the model structure. For example, the model presented in Section 2.2 factors as $f(x_1, x_2, t) = f_s(x_1, x_2) f_t(t)$ with the further assumption that $f_t(t)$ is quasi-periodic. To the best of our knowledge, encoding such structure into a tree-based model is not possible (Hastie et al., 2009, Ch. 9–10). In practical applications, various kinds of prior knowledge often exist, and the ability to utilize it in the analysis would improve the results. Therefore, GPs' ability to flexibly incorporate prior assumptions into the model structure makes them preferable to tree-based models in many cases. Additionally, as tree-based models are piecewise constant functions by definition, there is no way to control their smoothness (i.e. they are non-differentiable). Consequently, their fit is also typically jagged (see e.g. Elith et al., 2008) which is often undesirable, as species distributions are usually smooth in nature. GPs' ability to control the smoothness of the model fit thus makes them appealing in comparison to tree-based models when modelling species spatial or temporal (or spatio-temporal) distributions. Representing prediction uncertainty is also more challenging with tree-based models, although some estimates can be obtained with techniques such as bootstrapping (Hastie et al., 2009, Ch. 8). In science-based decision-making, a decision-maker often desires to know how confident one can be with the background information. Hence, GPs' well-calibrated uncertainty estimates make them an appealing choice in comparison to tree-based models in case studies such as ours, where the results will be used in political decision-making or management (see Section 4.4 for practical examples for suitable case studies).

4.3 | Future usage of Gaussian processes and the R package `gplite` in ecology

Our aim was to introduce and promote GP modelling as a powerful tool for analysing heterogeneous data and for revealing differences in the migration patterns of bean goose subspecies. Thereby, we only used time and location to predict the occurrence probability of taiga and tundra bean geese. For future reference, we emphasize that it is possible to include various environmental variables as covariates together with the model presented in this paper, and in that way study the biological factors behind the

phenomenon of interest. We also note that it is possible to apply GPs to model presence-only data with the point process modelling approach, for example using the log-Gaussian Cox process model (e.g. Diggle et al., 2013).

We also emphasize that GPs are not specifically designed for modelling data collected by citizens, but can also be implemented for other types of data (satellite tracking, geologging etc.). However, data collected by citizens from various taxa offer long and cost-efficient time series for ecological research from many parts of the world. These data provide under-utilized possibilities to study the spatio-temporal occurrence of animals. This study shows the feasibility of GP models for modelling citizen science data, and their capability to produce scientific knowledge for decision-making in management. In addition to this study, the management of the greylag goose *Anser anser* population in Europe is an example case where GP models could be used to improve management. An essential problem in the greylag goose case is to recognize when and where the migratory and sedentary parts of the population overlap (Bacon et al., 2019). The currently used method for distribution modelling (kernel density estimation, Bacon et al., 2019) does not provide any uncertainty estimation to the distributions, which would be achieved using GPs. Additionally, GPs would enable the construction of a quasi-periodic time component for modelling the distributions, which is an obvious assumption for distribution changes between years for most migratory birds. Together, these advances would make the results more transparent and, presumably, more accurate (see Section 4.4 for additional examples).

Furthermore, our study provides practical tools for implementing a variety of GP models (R package `gplite`). We point out that our software provides several additional features compared to the implementation in Golding and Purse (2016), which only allows for a Bernoulli observation model and a squared exponential kernel. The extra features in our R package `gplite` include several different covariance functions (e.g. neural network, Matérn, periodic) and a possibility to combine them, multiple observation models (Gaussian, binomial, beta-binomial, Poisson), sparse approximations for facilitating larger datasets and methods for model assessment and comparison.

4.4 | Management implications

Our results can be implemented not only to bean goose management at national and flyway levels, but also to the conservation and management of animals on a global scale. In the bean goose, the conservation of subspecies taiga bean goose is carried out at a flyway level, and harvest is managed internationally by applying an adaptive harvest management framework (Marjakangas et al., 2015). The hunting bag probably consists of both subspecies in many countries, but subspecies composition in the hunting bag is largely unknown (Heldbjerg et al., 2019). The legal hunting season for bean geese in Finland begins on 20 August and ends on 31 December, but the season can be shortened and the hunting area can be restricted geographically by the Ministry of Agriculture and Forestry of Finland. This kind of regulation is a common practice in harvest management in Finland, and our results provide a scientific base for adjusting the bean goose hunting season and area to meet the different management goals for both subspecies. The results from our case study show that bean goose harvest can be targeted at tundra bean goose in Finland by geographically restricting hunting to south-eastern Finland and by delaying the beginning of the hunting season from August to approximately the beginning of October. Naturally, our approach can also be used to predict the spatio-temporal distribution of bean goose subspecies also elsewhere in their range.

On a global scale, our approach combining citizen science data with GP modelling offers useful and cost-efficient predictions on spatio-temporal distributions of populations, which can be used to solve various management problems with animals from diverse taxa. For example, the ability of ticks to spread multiple zoonotic tick-borne diseases is known to vary between species, and sample collections by citizens have already been organized (Laaksonen et al., 2018). A combination of such data and GP modelling could enable finding spatio-temporal differences in the occurrences of various tick species, which could help to address vaccination campaigns more accurately. Similarly, our approach has obvious applications in fisheries: fishing is often targeted to multiple species or populations simultaneously, which makes the spatio-temporal regulation of fishing an important tool in sustainable fish stock management (Cooke et al., 2016). A great example of such a situation is the management of various river populations of Atlantic Salmon *Salmo salar* at the Baltic Sea. These populations return to their natal rivers annually for spawning, but spend the winters at sea. Targeting fishing at sea to the desired population is one of the key actions in the successful management of these salmon populations (Torniainen et al., 2014), and thus management would benefit from the knowledge of spatio-temporal differences in the occurrence of different river populations during winter. These differences could be studied using GP modelling together with citizen science data (such as stable isotopes of scales) or with professionally collected data (radiotracking, tag recovery). Finally, our approach could be used in the management of invasive species, where the management goal is to control or eradicate harmful populations while conserving other species. An excellent example of such a case is the introduced population of northern pike *Esox lucius* in south-central Alaska that is spreading and threatening the native salmonid populations (Dunker et al., 2020). As the pikes are controlled using extreme methods, such as poisoning the water systems, knowledge of the spatio-temporal occurrence of the pike population and other species exposed to the same management actions (salmonids, piscivorous birds, macroinvertebrates) is needed to minimize the negative effects of pike management. These patterns can be studied by combining GP modelling with suitable data from various populations.

ACKNOWLEDGEMENTS

The authors thank Birdlife Finland and regional ornithological societies for providing bird observation data, Stella Thompson for

reviewing the language and the Ministry of Agriculture and Forestry of Finland for providing a grant to T.L. and A.P.

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

A.P. conceived the original idea for the study, led the writing of the manuscript and obtained the data; J.P. designed and implemented the R package, carried out the analysis and wrote the methodological parts of the article; T.L. participated to the writing of the manuscript and supervised throughout the process. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

Data and example codes are available via GitHub https://github.com/jpiironen/anser_fabalis. For permanent link to the data and codes, see https://doi.org/10.5281/zenodo.5713729 (Piironen, 2021a). The publicly available data have been manipulated by adding a little bit of noise to the locations of all the observations. This will not affect the analyses, but was done according to the demand of the data owner (BirdLife Finland) to avoid any misuse of the data.

Installation instructions and a quick-start tutorial for the R package `gplite` can be found at https://github.com/jpiironen/gplite.

ORCID

Antti Piironen https://orcid.org/0000-0003-1986-9593
Juho Piironen https://orcid.org/0000-0002-0784-8835
Toni Laaksonen https://orcid.org/0000-0001-9035-7131

REFERENCES

Altwegg, R., & Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, *10*, 8–21.

Anderson, G. Q., & Green, R. E. (2009). The value of ringing for bird conservation. *Ringing and Migration*, *24*, 205–212.

Bacon, L., Madsen, J., Jensen, G. H., de Vries, L., Follestad, A., Koffijberg, K., Kruckenberg, H., Loonen, M., Månsson, J., Nilsson, L., Voslamber, B., & Guillemain, M. (2019). Spatio–temporal distribution of greylag goose *Anser anser* resightings on the north-west/south-west European flyway: Guidance for the delineation of transboundary management units. *Wildlife Biology*, *1*, wlb.00533.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hillb, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., & Stuart-Smith, J. F. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154.

Boettiger, C., Mangel, M., & Munch, S. (2015). Avoiding tipping points in fisheries management through Gaussian process dynamic programming. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20141631.

Bradter, U., Mair, L., Jönsson, M., Knape, J., Singer, A., & Snäll, T. (2018). Can opportunistically collected citizen science data fill a data gap for habitat suitability models of less common species? *Methods in Ecology and Evolution*, *9*, 1667–1678.

Callaghan, C. T., Rowley, J. J. L., Cornwell, W. K., Poore, A. G. B., & Major, R. E. (2019). Improving big citizen science data: Moving beyond haphazard sampling. *PLoS Biology*, *17*, e3000357.

Cheng, B. Y., Shyu, G. S., Wu, S. C., Lin, H. H., Hsu, C. H., LePage, B. A., & Fang, W. T. (2019). Fragmented riverine habitats in Taiwan have spatio-temporal consequences, re-distributing *Caprimulgus affinis* into urban areas leading to a human–wildlife conflict. *Sustainability*, *11*, 1778.

Cooke, S. J., Martins, E. G., Struthers, D. P., Gutowsky, L. F. G., Power, M., Doka, S. E., Dettmers, J. M., Crook, D. A., Lucas, M. C., Holbrook, C. M., & Krueger, C. C. (2016). A moving target – incorporating knowledge of the spatial ecology of fish into the assessment and management of freshwater fish populations. *Environmental Monitoring and Assessment*, *188*, 239.

Diggle, P. J., Moraga, P., Rowlingson, B., & Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Science*, *28*, 542–563.

Dunker, K., Massengill, R., Bradley, P., Jacobson, C., Swenson, N., Wizik, A., & DeCino, R. (2020). A decade in review: Alaska's adaptive management of an invasive apex predator. *Fishes*, *2*, 12.

Elith, J., Leathwick, J., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

Elith, J., Phillips, S. J., Hastie, T., Dud'ik, M., Chee, Y. E., & Yates, C. J. (2010). A statistical explanation of maxent for ecologists. *Diversity and Distributions*, *17*, 43–57.

Fox, A. D., & Madsen, J. (2017). Threatened species to super-abundance: The unexpected international implications of successful goose conservation. *Ambio*, *46*, 179–187.

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*, 598–608.

Halpern, B. S., Walbridge, S., Selkoe, K. A., Kappel, C. V., Micheli, F., D'Agrose, C., Bruno, J. F., Casey, K. S., Ebert, C., Fox, H. E., Fujita, R., Heinemann, D., Lenihan, H. S., Madin, E. M. P., Perry, M. T., Selig, E. R., Spalding, M., Steneck, R., & Watson, R. (2008). A global map of human impact on marine ecosystems. *Science*, *319*, 948–952.

Hastie, T., & Tibrishanhi, R. (1990). *Generalized additive models*. Chapman and Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer-Verlag.

Hawkins, A. S., Hanson, R. C., Nelson, H. K., & Reeves, H. M. (1984). *Flyways: Pioneering waterfowl management North America*. US Department of the Interior, Fish and Wildlife Service.

Heinicke, T. (2010). Tundra bean goose *Anser fabalis rossicus* during spring migration in northern Sweden – Rare visitor or regular passage migrant? *Ornis Svecica*, *20*, 174–183.

Heinicke, T. (2018). Western tundra bean goose *Anser fabalis rossicus*. In A. D. Fox & J. O. Leafloor (Eds.), *A global audit of the status and trends of Arctic and Northern Hemisphere goose populations (Component 2: Population accounts)*. Conservation of Arctic Flora and Fauna International Secretariat.

Heldbjerg, H., Fox, A.D., Christensen, T.K., Clausen, P., Kampe-Persson, H., Koffijberg, K., Liljebäck, N., Mitchell, C., Nilsson, L. & Skyllberg, U. (2019). *Taiga Bean Goose Population Status Report 2018–2019*. AEWA EGMP Technical Report Report No. **14**, Bonn, Germany.

Honka, J., Kvist, L., Heikkinen, M. E., Helle, P., Searle, J. B., & Aspi, J. (2017). Determining the subspecies composition of bean goose harvests in Finland using genetic methods. *European Journal of Wildlife Research*, *63*, 19.

Ingram, M., Vukcevic, D., & Golding, N. (2020). Multi-output Gaussian processes for species distribution modelling. *Methods in Ecology and Evolution*, *11*, 1587–1598.

Johnson, F. A., Alhainen, M., Fox, A. D., Madsen, J., & Guillemain, M. (2018). Making do with less: Must sparse data preclude informed harvest strategies for European waterbirds? *Ecological Applications*, *28*, 427–441.

Laaksonen, M., Klemola, T., Feuth, E., Sormunen, J. J., Puisto, A., Mäkelä, S., Penttinen, R., Ruohomäki, K., Hänninen, J., Sääksjärvi, I. E., Vuorinen, I., Sprong, H., Hytönen, J., & Vesterinen, E. J. (2018).

Tickborne pathogens in Finland: Comparison of *Ixodes ricinus* and *I. persulcatus* in sympatric and parapatric areas. *Parasites & Vectors*, *11*, 556.

Madsen, J., Bunnefeld, N., Nagy, S., Griffin, C., du Rau, P.D., Mondain-Monval, J.Y., Hearn, R., Grauer, A., Merkel, F.R. & Williams, J.H. (2015). *Guidelines on sustainable harvest of migratory waterbirds*. AEWA Conservation Guidelines No. 5, AEWA Technical Series No. 62, Bonn, Germany.

Madsen, J., Williams, J. H., Johnson, F. A., Tombre, I. M., Dereliev, S., & Kuijken, E. (2017). Implementation of the first adaptive management plan for a European migratory waterbird population: The case of the Svalbard pink-footed goose *Anser brachyrhynchus*. *Ambio*, *46*, 275–289.

Marjakangas, A., Alhainen, M., Fox, A.D., Heinicke, T., Madsen, J., Nilsson, L. & Rozenfeld, S. (2015). *International single species action plan for the conservation of the taiga bean goose* (*Anser fabalis fabalis*). AEWA Technical Series No. 56, Bonn, Germany.

McKinley, D. C., Miller-Rushing, A. J., Ballard, H. L., Bonney, R., Brown, H., Cook-Patton, S. C., Evans, D. M., French, R. A., Parrish, J. K., & Phillips, T. B. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, *208*, 15–28.

Nichols, J. D., Runge, M. C., Johnson, F. A., & Williams, B. K. (2007). Adaptive harvest management of north American waterfowl populations: A brief history and future prospect. *Journal of Ornithology*, *148*, S343–S349.

Nilsson, L. (2011). The migration of Finnish bean geese *Anser fabalis* in 1978–2011. *Ornis Svecica*, *21*, 157–166.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *40*, 1–42.

Patil, A. (2007). *Bayesian nonparametrics for inference of ecological dynamics* (PhD thesis), University of California, Santa Cruz.

Phillips, S., Anderson, R., & Schapire, R. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Piironen, J. (2021a) Data from: Predicting spatio-temporal distributions of migratory populations using Gaussian process modelling, *Zenodo*, https://doi.org/10.5281/zenodo.5713729

Piironen, J. (2021b). *gplite: Implementation for the most common Gaussian process models*. R package. Retrieved from https://github.com/jpiironen/gplite

Prasad, A., Iverson, L., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*, 181–199.

Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, *6*, 1939–1959.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Riutort-Mayol, G., Bürkner, P.C., Andersen, M.R., Solin, A. & Vehtari, A. (2020). Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *arXiv:2004.*11408.

Sheaffer, S. E., Rusch, D. H., Humburg, D. D., Lawrence, J. S., Zenner, G. G., Gillespie, M. M., Caswell, F. D., Wilds, S., & Yaich, S. C. (2004). Survival, movements, and harvest of eastern prairie population canada geese. *Wildlife Monographs*, *156*, 1–54.

Sigourney, D. B., Munch, S. B., & Letcher, B. H. (2012). Combining a Bayesian nonparametric method with a hierarchical framework to estimate individual and temporal variation in growth. *Ecological Modelling*, *247*, 125–134.

Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 1257–1264). MIT Press.

Tomkiewicz, S. M., Fuller, M. R., Kie, J. G., & Bates, K. K. (2010). Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 2163–2176.

Torniainen, J., Vuorinen, P. J., Jones, R. I., Keinänen, M., Palm, S., Vuori, K. A., & Kiljunen, M. (2014). Migratory connectivity of two baltic sea salmon populations: Retrospective analysis using stable isotopes of scales. *ICES Journal of Marine Sciences*, *71*, 336–344.

Vanhatalo, J., Hartmann, M., & Veneranta, L. (2020). Additive multivariate Gaussian processes for joint species distribution modeling with heterogeneous data. *Bayesian Analysis*, *15*, 415–447.

Vanhatalo, J., Veneranta, L., & Hudd, R. (2012). Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (*Coregonus lavaretus* l. s.l.) larvae. *Ecological Modelling*, *228*, 49–58.

Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Computation*, *10*, 1203–1216.

Williams, C. K. I., & Rasmussen, C. E. (1996). Gaussian processes for regression. *Advances in Neural Information Processing Systems*, *8*, 514–520.

Wright, W. J., Irvine, K. M., Rodhouse, T. J., & Litt, A. R. (2021). Spatial Gaussian processes improve multispecies occupancy models when range boundaries are uncertain and nonoverlapping. *Ecology and Evolution*, *11*, 8516–8527.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Piironen, A., Piironen, J., & Laaksonen, T. (2022). Predicting spatio-temporal distributions of migratory populations using Gaussian process modelling. *Journal of Applied Ecology*, 00, 1–11. https://doi.org/10.1111/1365-2664.14127