

Yihan Cao

Statistical methods for estimating fluctuating selection

Doctoral thesis
for the degree of Philosophiae Doctor

Trondheim, February 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU

Norwegian University of Science and Technology

Doctoral thesis
for the degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

© 2020 Yihan Cao. All rights reserved

ISBN (printed version)
ISBN (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU,

Printed by NTNU-trykk

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree Philosophiae Doctor (Ph.D) at the Department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU). The whole work was financially supported by the Research Council of Norway through its Centres of Excellence funding scheme (project number 223257 to Centre for Biodiversity Dynamics (CBD)).

I would like to express my gratitude to my main supervisor, Professor Jarle Tufto at NTNU, for allowing me to do Ph.D in Norway and for his tireless guidance on my papers and thesis. I also would like to thank my co-supervisor, Professor Marcel E. Visser at NIOO-KNAW, for his guidance, encouragement, and consideration along my way to Ph.D. I thank my office mates and colleagues around me at CBD for sharing me with their research and knowledge, which helps me a lot with the study on evolutionary biology. I am also grateful to the colleagues in the statistics group at the Department of Mathematical Sciences for the talks, discussions, seminars over these three years.

Finally, I would thank my friends in Trondheim for their company and support through times good and bad in daily life during these years. I also thank my family for their understanding of my choices, their spiritual support, and encouragement in everything I do in life.

Contents

Preface	i
List of papers	1
Introduction	2
Fluctuating and auto-correlated selection	3
Selection on correlated traits	4
Selection via multiple fitness components	5
Partial and complete brood failure	6
Ecological selective agents	6
Double brooding evolution	6
Bayesian analysis of ecological processes	7
Aims	9
Methods	11
The great tit study system	11
State-space models	11
Vector autoregression	12
Zero-inflated models	13
Laplace approximation to deal with random effects	14
MCMC sampling from a TMB model	15
The evolution of double brooding	16
Main results and discussion	17
Zero-inflated observations	17
Directional selection via complete brood failure	17
Stabilizing selection via expected number of fledglings (and offspring survival)	17
Ecological drivers of selection	18
Decreasing probability of double-brooding	19
Laplace approximation in <i>tmbstan</i>	19
Future perspectives	20
References	23
Paper I-IV	

List of papers

- I **Y. Cao**, M. E. Visser, and J. Tufto, 2019. A time-series model for estimating temporal variation in phenotypic selection on laying dates in a dutch great tit population. *Methods in Ecology and Evolution* 10(9), 1401–1411.
- II J. Tufto, **Y. Cao**, M. E. Visser. Evolution of double brooding. In revision.
- III **Y. Cao**, M. E. Visser, and J. Tufto. Multi-episodic fluctuating selection via fertility and viability in a great tit (*Parus major*) population. Manuscript.
- IV **Y. Cao**, M. E. Visser, and J. Tufto. Bayesian inference with tmbsstan for a state-space model with VAR(1) state equation. Manuscript.

Declaration of contributions

Paper I: M.E.V. provided the data; J.T. conceived the idea and initiated the statistical model; Y.C. analyzed the data and conducted the analyses; Y.C. wrote the initial draft with input from J.T.; all authors contributed to revisions on later manuscript versions and gave final approval for publication.

Paper II: J.T. conceived the idea and initiated the basic genetic model; J.T. derived analytical expression of the genetic model under different scenarios with input from Y.C.; J.T. conducted the numerical analyses with the theoretical models; M.E.V. provided the data for empirical study; J.T. and Y.C. conducted the empirical study with the data; J.T. wrote the initial draft with input from Y.C.; the final version was reviewed and commented by all co-authors.

Paper III: M.E.V. provided the data; J.T. conceived the idea; Y.C. initiated the statistical model, analyzed the data and conducted the analyses; Y.C. wrote the initial draft with input from J.T.; all authors contributed to revisions on later manuscript versions and gave final approval for publication.

Paper IV: J.T. conceived the idea; Y.C. set up the simulation studies with input from J.T.; Y.C. conducted simulations and analyzed simulation results; Y.C. conducted the empirical study with the data provided by M.E.V.; Y.C. wrote the manuscript and it was reviewed and commented by all co-authors.

All authors gave final consent for the use of listed work above in this thesis.

Introduction

Natural selection is a key mechanism of evolution and the central process in nature. It occurs when there is a difference between phenotypical traits in expected relative fitness (Gardner and Grafen, 2009; Hansen, 2017). It also plays a role in shaping life cycles in ways that optimize reproductive fitness (Charnov, 1993; Stearns, 2000) and the mechanism of which has been studied in the framework of life history theory. In population biology, one of the fundamental questions is which selection, under which circumstances and to what extent, can have an appreciable impact on population dynamics (Charlesworth, 1971; Saccheri and Hanski, 2006). Understanding the genetic basis of the traits that selection operates on and the signatures of past and present selection in patterns of variation in the genome remain as a priority in the research agenda for evolutionary biologists (Stinchcombe et al., 2017). Even though the interplay between selection and life history evolution, selection and population dynamics has been approached from various perspectives in each study area over the past years, to obtain a better understanding of the role of natural selection in driving evolutionary changes, accurate estimates of the strength of selection acting in the wild is an essential prerequisite (Linnen and Hoekstra, 2009).

Most of the previous work attempting to measure natural selection within populations drew on the seminal studies of Price (1970); Lande (1979); Lande and Arnold (1983), in which the selection is characterized by the relationship between traits and relative fitness. Building on their foundational work, Schlüter (1988) provides a non-parametric estimate of the fitness function and uses it to suggest an appropriate parametric model. Thomson and Hadfield (2017) shows that using offspring fitness components as part of parental fitness (“mixed fitness” in their terms) is common in studies of birds and mammals, but will only lead to correct estimates of selection and evolutionary change under very restrictive conditions. Among many others, the enormous literature contains conceptual, methodological and statistical recommendations to estimating the phenotypic covariances between traits and some aspect of relative fitness (Stinchcombe et al., 2017). In empirical studies, the mode and intensity of natural selection are estimated by regressing relative fitness onto phenotypic values. The selection gradient (β) analysis has now been applied to a wide range of plant and animal taxa (reviewed by Kingsolver et al., 2001; Siepielski et al., 2009).

Fluctuating and auto-correlated selection

The publication of synthetic reviews of form and strength of selection (Kingsolver et al., 2001; Siepielski et al., 2009) confirms that phenotypic selection commonly fluctuates in strength and frequently changes in direction among years. The variance in phenotypic selection was usually estimated by computing the variance of the strength of selection using selection gradients estimated separately at each time point which reflects both random sampling error and real variation in selection (Morrissey and Hadfield, 2012). Since temporal variation in natural selection is a fundamental determinant of evolutionary outcomes and an appealing hypothesis to explain evolutionary stasis (Price and Liou, 1989; Merilä et al., 2001; Siepielski et al., 2010), more accurate models with a detailed look at the extent of variation in selection, accounting for sampling error, are desirable. Among the previous empirical studies accounting for the sampling error of variation, Calsbeek (2011) presents a non-parametric analysis in exploring the variation of fitness

surfaces over time or space. In contrast, using a log-quadratic generalized linear mixed model with a random effect on the regression slope implemented using Integrated Nested Laplace Approximations (INLA, Rue et al., 2009), Chevin et al. (2015) estimated yearly fluctuations and autocorrelation in optima of a Gaussian fitness function. Using instead the more flexible framework of Template Model Builder (TMB, Kristensen et al., 2016), Gamelon et al. (2018) fitted a model of fluctuating selection via several non-overlapping selection episodes with non-linear random effects added directly on the location of the fitness optima and on the peak of the fitness function. The latter two identify the pattern of temporal dynamics in the selection not only by its variance but also by its temporal auto-correlation. Previous theory has shown that the auto-correlation of selection strongly affects whether (and how much) genetic responses to selection optimize long-term fitness and population growth in a fluctuating environment (Charlesworth, 1993; Lande and Shannon, 1996; Bürger and Gimelfarb, 2002; Chevin, 2013; Tufto, 2015). The empirical estimate of auto-correlation in the location of the fitness optima turned out to be significant in Chevin et al. (2015); Gamelon et al. (2018); Cao et al. (2019). The generality of this finding, however, needs to be confirmed across a wider range of species, populations, and traits, using the same, statistically robust approach. As of yet, estimating auto-correlation in selection may require a large sample size with many time points (Chevin and Haller, 2014).

A straightforward extension of previous models with temporally varying selection strength for stabilizing selection is to allow all the properties (height, location of maximum and width) of a Gaussian fitness function at population level to be temporally fluctuating and even cross-correlated. Such a statistical model including all these possibilities into one framework can be as complex as that powerful enough model-fitting techniques are required for statistical inference. Cao et al. (2019) is among the few to have done this with an R package named Template Model Builder (TMB, Kristensen et al., 2016), which is developed for fast-fitting complex, linear or nonlinear statistical models. The temporal fluctuation in the strength and even the direction of selection can be captured by using appropriate statistical approaches. However, changes in the form of selection, which are likely common, are harder to quantify (Siepielski et al., 2009).

Selection on correlated traits

The target trait that selection acts on can be correlated with fitness either because they impact fitness directly (direct selection) or because they are correlated with other traits that affect fitness (indirect selection) (Linnen and Hoekstra, 2009). For the great tits, the brood size is correlated with the egg-laying date and the early breeders tend to lay bigger clutches (Perrins and McCleery, 1989; Barba et al., 1995). In a black-throated blue warbler population, the egg-laying date of the first brood is positively correlated with the propensity a second brood to be laid from a given female (Townsend et al., 2013). We tend to focus on traits that we have a priori reasons to believe are targets of selection. In fact, strong indirect selection can overcome direct selection in an opposing direction (Linnen and Hoekstra, 2009). How can we determine the actual target of selection? Lande and Arnold (1983) shows elegantly how total selection can be partitioned into direct selection on a trait and indirect selection through correlated traits, in which selection gradients (β) are calculated using multiple regression to control for indirect selection, thereby estimating direct selection on a trait. The famous Darwin's finches also illustrate the importance

of measuring multiple traits and estimating both direct and total selection.

The correlated characters that selection is acting simultaneously on might likely be genetically correlated, so selection on one trait can result in a change in the other. The total response to selection will be a combination of direct selection on a particular trait, plus indirect selection resulting from a correlated response to selection on some other traits, and therefore leads to an accurate picture of adaptation and evolutionary constraint in natural populations. In reality, the data on genetic structure of correlated traits are not always available, it is thus necessary to conduct simulations with various genetic and phenotypic covariance structures for correlated traits, explore the evolutionary trajectories under different scenarios and compare them with the reality, to gain a better understanding of the mechanism behind the correlational selection on the traits. Alternatively, Reed et al. (2016) uses an animal model to obtain the genetic covariance matrix of clutch size and laying date and then calculates predicted response to selection based on the Robertson–Price Identity and the multivariate breeder’s equation (MVBE). It finally concludes that the similar prediction indicates that unmeasured covarying traits were not missing from the analysis.

Selection via multiple fitness components

Most studies estimating natural selection focus on a specific component. For short-lived hole-nested species, pre-breeding mortality is one of the major sources of individual variation in lifetime reproductive success (Clutton-Brock, 1988; Newton et al., 1989), which implies that the fate of individual fledglings is completely altered after recruiting to the population. This phenomenon can be recognized as a straightforward reason of different selection patterns estimated with the same populations since either number of fledglings or recruits is taken as the fitness component (for example Verboven and Visser, 1998; Reed et al., 2013a; Chevin et al., 2015), but rarely both (except for Gamelon et al., 2018). How the temporal dynamics of phenotypic selection may vary among fitness components (e.g. fecundity and survival) is poorly understood thus (Siepielski et al., 2010). Furthermore, many previous studies (for example Siikamäki, 1998; Verboven and Visser, 1998) have demonstrated that the date of fledgling affects post-fledgling survival, the usual pattern being early fledglings experienced higher survival. An advancement of mean annual egg-laying date is thus expected to be observed to maximize offspring fitness, however, the reality contradicting the expectation is that an enlarging mistiming between the egg-laying date and food peak date over the course of study is observed (Visser et al., 1998; Chevin et al., 2015; Cao et al., 2019). One potential explanation is that the adaptive evolutionary change is determined by relative form and strength of selection acting among different fitness components (Schluter et al., 1991; Hoekstra et al., 2001). Besides, integrating multiple fitness components into one modeling framework is a start point to explore the evolution of life history traits (e.g. size at birth, number, size, and sex of offspring, lifespan) and the dynamic interaction between them, which is research objectives in life history theory.

Even though the importance of measuring selection through separate episodes of selection over the reproductive cycle was pointed out by Arnold and Wade (1984), the empirical measurements on selection have rarely done this. The exceptions include Engen et al. (2011), in which selection is estimated separately with fitness components (fecundity and survival) in different

age classes. Gamelon et al. (2018) proposes a multi-episodic approach where different reproductive stages (clutch size, survival from egg to fledgling, from fledgling to recruit and breeding mothers) are included in one statistical model. Potential ecological drivers of selection on both laying dates and clutch sizes were accounted and the method was applied to a dipper population.

Partial and complete brood failure

In altricial birds, the nestlings are brooded for 1 to 2 weeks after hatching and typically obtain extensive parental care from both parents before independence (Liker et al., 2015). Partial and complete brood failure is common in this period and this is a key determinant of variation in reproductive success in such species (Santema and Kempenaers, 2018). The underlying causes of nestling mortality are usually unknown unless the nest predation is identified (Martin and Briskie, 2009). In some bird species, complete brood failure is found associated with nest predation, (McCleery et al., 1996) which might be related to nest-site security (Wesołowski, 2002) and to the sudden and permanent disappearance of one of the parents (Santema and Kempenaers, 2018). It is often hypothesized that offspring mortality results from a particular factor such as breeding timing that determines brood success through its effect again on parental care. Even though it is plausible that a particular factor influences both partial and complete brood mortality, the effect sizes of the factor on them likely differ. Moreover, if partial and complete brood mortality has different proximate causes, it might give misleading results on the effects of biological factors on offspring mortality when they are lumped together. Therefore, it is biologically and statistically necessary to separate complete brood failure from partial brood failure when exploring the proximate mechanism of offspring mortality.

Ecological selective agents

Changes in ecological conditions driven by climatic fluctuations appear to be common and important. Natural selection on wild populations is driven by such changes in biotic and abiotic conditions (Bell, 2010). Despite of increasing interests in the environmental sensitivity of phenotypic selection, few studies have identified causal mechanisms underlying temporal variation in the form, direction, and strength of selection (Siepielski et al., 2009). Several studies have linked temporal variation in natural selection through survival or fecundity to variation in ecological factors such as density, temperature, precipitation, predation, competition, and many other factors. These factors are heterogeneous at both temporal and spatial scales. For example, the survival of juveniles is identified to be strongly density-dependent (Reed et al., 2013a,b) and density is shown to be a varying selective agent in a dutch great tit population (Sæther et al., 2016). The temporal variation in optimal phenotypic maximizing yearly fitness subjects to fluctuating spring temperature (Chevin et al., 2015; Gamelon et al., 2018). Predation is a selective pressure leading to fledglings hatched early in the season suffering lower probability of complete brood failure in great tits (Sæther and Bakke, 2000). In turn, the changing climate conditions lead phenotypic distribution to be constantly shaped and reshaped by various agents of natural selection (Endler, 1986). Even though these studies have accumulated our understanding of environmental sensitivity in natural selection, incorporating abiotic and biotic factors as potential selective agents into the big picture of estimating varying selective selection on various traits throughout the life cycle has remained challenging.

Double brooding evolution

Multiple breeding (more than one reproductive attempt in a breeding season) is a common reproductive strategy in short-lived species (Verhulst et al. 1997 and references therein). The frequency of double brooding is an important factor determining the productivity of a population, as Nagy and Holmes (2004) shows that 19% of the annual variance in fecundity is explained by double brooding in a black-throated blue warbler population in America. Since annual fecundity plays a major role in determining population growth (Sæther and Bakke, 2000), understanding the mechanism of multi-brooding in short-lived species has implications on the future viability of a population. Several studies of birds have investigated the intra-seasonal costs (Mulvihill et al., 2009) or determinants (Jacobs et al., 2013) of multiple-breeding, either experimentally (Parejo and Danchin, 2006) or using longitudinal studies (Townsend et al., 2013) or combination of them (Evans Ogden and Stutchbury, 1996; Verboven and Verhulst, 1996). These studies find that delaying hatching date, as well as increasing clutch size and/or brood size, commonly lead to a lower probability of initiating a second clutch (Lindén, 1988; Geupel and DeSante, 1990; Evans Ogden and Stutchbury, 1996; Verboven and Verhulst, 1996; Verboven et al., 2001; Parejo and Danchin, 2006; Townsend et al., 2013). The study species include wren tit, hooded warbler, black-throated blue warbler, great tit, and many others. Husby et al. (2009) shows that in four long-term study populations of great tits in the Netherlands, the proportion of females that double brood has declined in all populations. They stated that the decline has two-fold reasons. The first is the increase in the mistime to the food peak experienced by the population over the study years and thus birds are less likely to attempt a second clutch. The second is the temporal decline in the number of recruits produced from the second clutch. They concluded that changing environmental conditions are important in determining the number of clutches a female lays and therefore potentially alter important life-history traits in the species.

These studies no doubt give us a better understanding of the mechanism of multiple brooding and provide promising explanations for the observational temporal fluctuations in the frequency of double-brooding. However, little theoretical and mechanical hypotheses for the double-brooding evolution exists. It is unclear if there is a genetic basis of the liability of multiple brooding and how the genetic structure interacts with different climate scenarios to produce different evolutionary consequences of double brooding. Due to the lack of genetic data on these reproductive traits of natural bird populations, investigating the mechanism of double brooding evolution is probably feasible only through theoretical genetic models.

Bayesian analysis of ecological processes

Both frequentist and Bayesian inferences are powerful tools for a better understanding of ecological processes in population and community ecology. In the frequentist framework, the most state-of-the-art model fitting technique, an R package named Template Model Builder (TMB, Kristensen et al., 2016) is gaining popularity recently due to its power and efficiency in fitting complex nonlinear mixed models, which are common when modelling complicated ecological processes (for example Cadigan, 2015; Albertsen et al., 2016; Auger-Méthé et al., 2017). One worth mentioning feature of TMB is that it enables Laplace approximation of the marginal likelihood where the random effects are automatically integrated out. Maximum marginal like-

lihood estimation with the Laplace approximation tends to be orders of magnitude faster but potentially leads to biased inference (Monnahan and Kristensen, 2018). In spite of the flexibility and efficiency of TMB, however, the lack of capability of working in the Bayesian framework has hindered the adoption of it for Bayesians. In the Bayesian framework, Bayesian statistical inference is used extensively to model dynamics of single species, population dispersal, growth, and extinction (Ellison, 2004). The software package *Stan* (Gelman et al., 2015), a probabilistic programming language for Bayesian statistical inference written in C++ is attracting people's attention in many fields as an alternative to BUGS (Lunn et al., 2000) and recommended to be widely applied in ecology due to its improved efficiency (Monnahan et al., 2017).

To best utilize the merits of both TMB and Stan, a new R package *tmbstan* (Kristensen, 2018) was developed to allow users to make Bayesian statistical analysis with TMB models. It provides MCMC sampling for TMB models while the integration of randoms effects can be calculated either with Laplace Approximation (by specifying *laplace=TRUE*) or with Stan. Monnahan and Kristensen (2018) conducts simulation studies and real case studies to compare the computational efficiency of *tmbstan* with and without Laplace approximation and check the validity of Laplace approximation. They found that enabling the Laplace approximation was less efficient than full MCMC integration, but it is unclear whether this will typically be true. The case studies also showed the Laplace approximation is not always met. Even though it is intuitive to apply *tmbstan* to estimating fluctuating natural selection especially when prior knowledge on some parameters is available, this has not been done to date. Therefore, there exists no guideline on whether Laplace approximation should be used to achieve better efficiency especially when the statistical model for estimating selection is extremely complicated. To answer this question, simulation studies under different scenarios in different statistical frameworks are necessary.

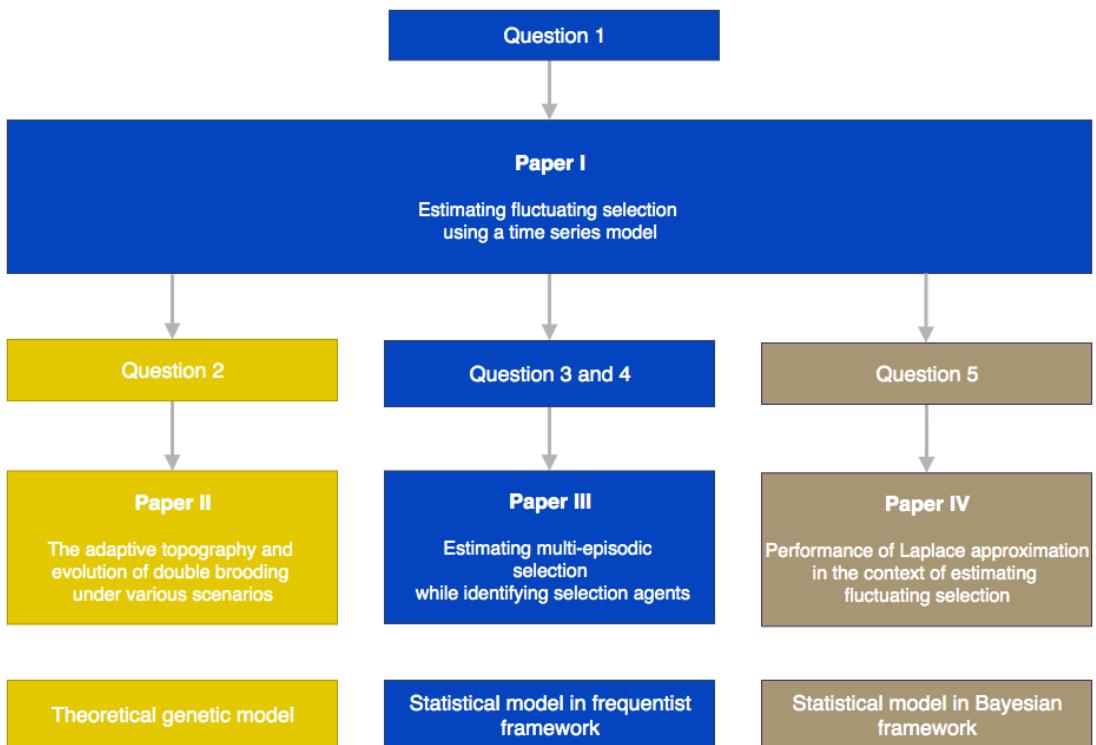
Aims

Linking the sources of natural selection to the dynamics of evolution has been a major goal of evolutionary biology, however, the lack of a unified framework to quantify the fluctuations in selection accurately has hampered this progress. Previous empirical findings show that fitness landscapes are not constant over time, and populations are evolving towards a continuously changing fitness optimum. A more statistically robust approach, however, is needed to be applied to a wider range of species, populations, and traits. This thesis contributes to this end by showing how current methods for estimating fluctuations in selection can be extended using a more flexible statistical framework. Due to the flexibility of the method equipped with state-space models and TMB, it can be extended to estimating fluctuating selection of life history for different life cycle segments while identifying biotic and abiotic factors exerting selective pressures and identifying which traits (egg-laying date or clutch size in our study), or combinations of traits (potentially correlated), will be targets of the selection. The method can be alternatively implemented in the Bayesian framework by taking prior information into account and using the Bayesian inference tool *tmbstan*. Using long-term brood-based data from a great tit population in the Netherlands, we hope to be able to answer the questions below:

- 1. Is there temporal variation, auto-correlation, and cross-correlation in phenotypic selection on the egg-laying date? (paper I, III)**
- 2. What is the possible explanation for the observed decline in the frequency of double brooding? (paper I, II)**
- 3. How selection operates on phenotypes differently in different selective episodes? (paper III)**
- 4. Which ecological variables drive the temporal variation in the phenotypic selection? (paper III)**
- 5. Is Bayesian inference made by "tmbstan" comparable with frequentist inference for estimating phenotypic selection and should Laplace approximation be used? (paper IV)**

The diagram in Fig. 1 shows the connection and transition of the papers in my dissertation. To be specific, paper II, III and IV are extended from paper I by asking specific questions listed above that are not addressed in paper I. According to the modeling approach used in the study, paper I, III and IV are grouped into "statistical model" and for paper II, it is "theoretical genetic model". Furthermore, Paper I and III are classified in the frequentist framework, while paper IV in the Bayesian framework, as illustrated by different colors of the blocks in the last row of the diagram.

Figure 1: A diagram showing how the papers in this dissertation are connected. Paper II, III and IV are extended from paper I by asking specific questions listed above. Generally, paper I, III and IV involve statistical modeling approaches and paper II theoretical genetic modeling approach, as illustrated by the last row. Furthermore, the studies in paper I and paper III were carried out in the frequentist framework, while paper IV in the Bayesian framework, as indicated by the different colors of the blocks.



Methods

The central elements of our statistical methods are great tit data (paper I, II, III and IV), state-space models (paper I, III, and IV), vector autoregression (paper I, III, and IV), zero-inflated models (paper I, III), Template Model Builder (paper I, III, and IV), tmbstan (paper IV), and evolution of double brooding (paper II).

The great tit study system



Figure 2: Map of the park (National Park of Hoge Veluwe in the Netherlands) where the great tit data have been collected.

Figure 3: A great tit.

The great tit (*Parus major*, Fig 3) is 18-20g small passerine bird species widespread throughout European woodlands and gardens. As a cavity nester, it readily accepts nest-boxes for breeding, which allows monitoring of the whole population if a surplus of nest-boxes is provided (Harvey et al., 1979). The study area ($52^{\circ}02' - 52^{\circ}07'N$, $5^{\circ}51' - 5^{\circ}32'E$ in The Netherlands, Fig 2) consists of mixed pine-deciduous woodland on poor sandy soils. From 1955 to 2015, more nest boxes than needed were placed in the study area at approximately constant availability. On average the ratio of nest boxes to breeding females was around 3:1 in a typical year. A surplus of nest boxes is supplied so that the actual number of individuals that survive is generally determined by selection and not by external limiting factors such as the number of nest sites. During the breeding season from April to June/July, nest boxes were visited once per week. At each visit, the number of eggs or nestlings was counted and nestlings were given metal leg rings on day 7 and the parents caught on the nest using a spring trap. For some years, clutch or brood size manipulation experiments were carried out, which possibly affected fledgling production or recruitment probability, therefore, manipulated broods were excluded from our studies.

State-space models

A State-Space Model (SSM) is a time series model where observations are regarded as made up of distinct components such as trend, seasonal, regression elements and disturbance terms (Durbin and Koopman, 2012). A typical SSM consists of two equations:

$$\mathbf{x}_t = g(\mathbf{x}_{t-1}, \mathbf{c}); \quad (1)$$

$$\mathbf{y}_t = h(\mathbf{x}_t, \nu), \quad (2)$$

where equation (1) is a process model describing the relationship between unobserved states with function g and parameter \mathbf{c} and \mathbf{x}_t denotes the state at time t . The observation model in equation (2) links the observation or measurement \mathbf{y}_t with state \mathbf{x}_t at time t through function h and parameter ν .

State-space models are often used for analyzing complex ecological processes that can not be observed directly, such as marine animal movement (Albertsen et al., 2015), population dynamics (Wang, 2007) and animal behavior (Morales et al., 2004). It provides a natural paradigm for ecosystem modeling (Pedersen et al., 2011). In spite of the flexibility of SSMs for estimating the unobserved states while simultaneously relating them to various environmental (and other) covariates of interest, these models and their implementations still have limitations arising from underlying difficulties of likelihood computation and maximization for non-Gaussian and non-linear models. Johnson et al. (2008) utilized the computationally efficient Kalman filter to compute the model likelihood but it is applicable only to linear Gaussian SSM formulations. Jonsen et al. (2005) and McClintock et al. (2012) relied on Markov Chain Monte Carlo (MCMC) techniques performed by sampling from the posterior likelihood of the parameters and the unobserved states, but it is computationally expensive and comparatively slow. Pedersen et al. (2011) examines and compares the estimation performance of three methods for fit of a theta logistic model for population dynamics with simulated data, namely Hidden Markov Model (HMM), AD Model Builder (ADMB) and the popular Bayesian framework of BUGS. It concludes that estimation performance for all three methods are largely identical, while ADMB establishes computing time superiority. The most state-of-the-art statistical tool named Template Model Builder (TMB) that can be used for fitting state-space models will be introduced later.

Vector autoregression

Vector autoregression (VAR) is a stochastic process model used to capture the linear interdependencies among multiple time series. It is an extension of the univariate autoregression model to multivariate time series data and consists of a list of models that can be hypothesized to affect each other intertemporally. All variables in a VAR enter the model in the same way: each variable has an equation explaining its evolution based on its own lagged values, the lagged values of the other model variables, and an error term.

The basic p -lag vector autoregressive (VAR(p)) model has the form:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{e}_t, t = 1, \dots, T, \quad (3)$$

where each \mathbf{y}_i is a vector of length k , each \mathbf{A}_i is a $k \times k$ coefficient matrix and \mathbf{e}_i is a $k \times 1$ unobservable zero mean white noise vector. Here I write a first-order VAR (VAR(1)) in a large matrix notation as

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} + \begin{bmatrix} a_{1,1}^1 & a_{1,2}^1 & \cdots & a_{1,k}^1 \\ a_{2,1}^1 & a_{2,2}^1 & \cdots & a_{2,k}^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1}^1 & a_{k,2}^1 & \cdots & a_{k,k}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{k,t} \end{bmatrix}. \quad (4)$$

Each variable $(y_{1,t}, y_{2,t}, \dots, y_{k,t})$ in the model has one equation. The current (time t) observation of each variable depends on its own lagged values as well as on the lagged values of each other variable in the VAR(1). Vector (c_1, c_2, \dots, c_k) is a k -vector of constants (intercepts). The matrix consisting of $a_{1,1}^1$ and so on is called transition matrix or autoregressive matrix. Vector $(e_{1,t}, e_{2,t}, \dots, e_{k,t})$ is errors that are usually assumed to be multivariate normal distributed. Variables $(y_{1,t}, y_{2,t}, \dots, y_{k,t})$ are cross-correlated either through the transition matrix or variance-covariance matrix of $(e_{1,t}, e_{2,t}, \dots, e_{k,t})$. To guarantee this VAR(1) process to be stationary, it is sufficient to ensure that the eigenvalues of the transition matrix lie in unit circle (Lütkepohl, 2005; Wei, 2006).

Zero-inflated models

In ecological research, most count data are zero-inflated. In our analyzed data set, for example, the response variable (number of chicks, number of fledglings, number of recruits) contain more zeros than expected based on the Poisson or negative binomial distribution. A zero-inflated model is a statistical model based on a zero-inflated probability distribution that can deal with the excessive number of zeros. The common used zero-inflated models for count data include zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-altered Poisson (ZAP) and zero-altered negative binomial (ZANB) models. The latter two differ from ZIP and ZINB in terms of the nature of zeros. ZAP and ZANB are two-part models that can deal with false zeros (see Zuur et al. (2009) for the definition of false zeros). The negative binomial models (ZINB and ZANB) can cope with overdispersion not only due to excessive numbers of zeros but also due to extra variation in the count data. The main R packages for modeling zero-inflated data include pscl, INLA, MCMCglmm, glmmADMB, mgcv, brms, gamm4 and glmmTMB (Zeileis et al., 2008; Rue et al., 2009; Hadfield et al., 2010; Skaug et al., 2013; Wood et al., 2016; Bürkner et al., 2017; Stasinopoulos et al., 2017; Magnusson et al., 2017). Brooks et al. (2017) makes a comparison between the packages and claims that glmmTMB is most appealing to users in terms of the combination of speed and flexibility.

In our analysis, zero-inflated Poisson (ZIP) model was used to deal with the excessive number of zeros in the number of fledglings and zero-inflated Beta-Binomial (ZIBB) was used to model the offspring viability (in our study offspring viability is defined as non-zero inflation probability \times offspring survival probability), in which there are excess zeros and the remaining component (offspring survival probability) can be modeled with predictors instead of being a fixed parameter. To be specific, a ZIP model consists of two components (equations (5) and (6)) corresponding to two zero generating processes. The first is governed by a binary distri-

bution and second by a Poisson distribution, which also generates zero counts. The two model components are described as follows with probability mass functions f :

$$f(y = 0) = \pi + (1 - \pi)e^{-\lambda}; \quad (5)$$

$$f(y|y \geq 1) = (1 - \pi)\frac{\lambda^y e^{-\lambda}}{y}, \quad (6)$$

where the outcome variable y has any non-negative integer value. The expected Poisson count is denoted as λ and π is the probability of extra zeros.

A ZIBB model (see Hu et al. (2018) for more details) also consists of two zero-generating processes. One is again governed by a binary distribution and the other one by a Beta-binomial distribution in which the probability p is a random variable drawn from a beta distribution parameterized by α and β . The two components are given below:

$$f(y = 0) = \pi + (1 - \pi)f_{\text{beta-bino}}(0|n, \alpha, \beta); \quad (7)$$

$$f(y|y \geq 1) = (1 - \pi)f_{\text{beta-bino}}(y|n, \alpha, \beta), \quad (8)$$

where n is the total number of events with any non-negative integer value and y is the number of successes. π is again the probability of extra zeros. The probability mass function of a Beta-binomial distribution $f_{\text{beta-bino}}$ is given by:

$$f_{\text{beta-bino}}(y | n, \alpha, \beta) = \binom{n}{y} \frac{B(y + \alpha, n - y + \beta)}{B(\alpha, \beta)}, \quad (9)$$

which consists of a binomial function and a beta function B .

Laplace approximation to deal with random effects

Consider a hierarchical model where the data y depend on a parameter vector θ and random effects u , then maximum likelihood inference requires maximization of

$$L(\theta) = P_\theta(y) = \int P_\theta(y | u)P_\theta(u)du. \quad (10)$$

The evaluation of this integral proves often difficult. Various numerical or analytical approaches were proposed to calculate the approximation of the integral. Among them, Laplace's method has been widely used to approximate likelihoods (Raudenbush et al., 2000). In standard Laplace approximation, the natural log of the integrand is expanded in a second-order Taylor series and higher order terms diminish with big sample size, the approximation to the likelihood is thus given as

$$L^*(\theta) \propto \det(|H(\theta)|)^{-\frac{1}{2}} \times P_\theta \{y | \hat{u}(\theta)\} P_\theta \{\hat{u}(\theta)\}, \quad (11)$$

where

$$\hat{u}(\theta) = \operatorname{argmax}_u P_\theta(y)P_\theta(u)$$

and

$$H(\theta) = \frac{\partial^2 L}{\partial u^2} \ln \{P_\theta(y) P_\theta(\hat{u}(\theta))\}$$

(see Kristensen et al. (2016) for the review of Laplace approximation).

I mentioned glmmTMB as an R package that can deal with zero-inflated models. In our analyses, however, I used another R package named Template Model Builder (TMB, Kristensen et al., 2016) instead of glmmTMB as a model fitting technique to benefit from the flexibility in model formulation in TMB. The relationship between them is that glmmTMB is built on TMB and provides a user-friendly interface similar to lme4 for researchers who have difficulties with TMB since TMB requires users to formulate models with C++. The core feature of TMB is that it evaluates the integral with Laplace approximation. The procedure of using TMB to fit a statistical model can be summarized into three steps. Firstly, the joint likelihood for the data, the fixed effects, and the random effects are defined by the user as a C++ template function. Then the package evaluates and maximizes the Laplace approximation of the marginal likelihood where the random effects are automatically integrated out. This approximation and calculation of its derivatives are achieved by using reverse-mode automatic differentiation (up to order three) of the joint likelihood. At last, the approximated likelihood function and its derivatives are passed to optimizers in R such as nlmrb and optim.

The combination of reverse-mode automatic differentiation and Laplace approximation for high-dimension integrals allows for the efficient fitting of complex (nonlinear, non-Gaussian, and hierarchical) models with large multivariate data sets to perform parameter estimation (Fournier et al., 2012). The performance of TMB is superior to ADMB (Kristensen et al., 2016) and thus is gaining researchers to use it instead of ADMB to fit state-space models (for example Albertsen et al., 2015; Cadigan, 2015; Albertsen et al., 2016; Berg and Nielsen, 2016). Another model fitting tool that uses the Laplace approximation and is known to be computationally efficient is INLA (Rue et al., 2009), but it is restricted to fit a class of models where the random effects are Gauss-Markov random fields (Kristensen et al., 2016).

MCMC sampling from a TMB model

I mentioned in the introduction that `tmbstan` (Kristensen, 2018) as an R package developed for MCMC Sampling from TMB model objects using Stan (Team, 2017; Carpenter et al., 2017), is able to make efficient Markov chain Monte Carlo (MCMC) sampling for a broad range of Bayesian models. It is worth noting that `tmbstan` not only provides TMB users with a possibility of making Bayesian statistical analysis with Stan, but also takes advantage of the features of both TMB and Stan by utilizing the flexibility of TMB in the model specification as well as the high computational efficiency of Stan.

I have introduced that TMB uses the Laplace approximation to integrate random effects. However, Laplace approximation is not always accurate especially when the random effects u are not Gaussian distributed. In addition, the higher-order terms in the Taylor series not necessarily diminish as sample size increases in some special model classes (Raudenbush et al., 2000). In a Bayesian analysis, MCMC integrates all parameters and this allows us to check the accuracy of Laplace approximation in TMB. `tmbstan` is featured with an argument 'laplace'. When this

argument is enabled, TMB would integrate random effects and Stan integrates the rest fixed effects. The accuracy of the Laplace approximation thus can be tested by comparing the posterior distributions of the fixed effects with and without Laplace approximation enabled in `tmbstan` (Monnahan and Kristensen, 2018).

The evolution of double brooding

For a great tit population, consider reproductive traits z_1 and z_2 , for example, z_1 is the laying date of first brood and z_2 is the liability of initializing a second brood. The phenotypic values z_1 and z_2 are assumed to be jointly multivariate normal. I also assume the genetic and phenotypic variance-covariance matrix \mathbf{G} and \mathbf{P} of z_1 and z_2 , as well as the age-specific fecundity and mortality rates for each phenotype, remain nearly constant for a few generations.

With above assumptions and let \bar{z}_1 and \bar{z}_2 be the mean phenotypic values in a given generation, then the change in mean phenotypic values from one generation to the next is given by

$$\Delta \bar{\mathbf{z}} = \mathbf{G} \nabla \ln \bar{w}(\bar{z}_1, \bar{z}_2), \quad (12)$$

where $\nabla = (\frac{\partial}{\partial \bar{z}_1}, \frac{\partial}{\partial \bar{z}_2})^\top$ is the gradient operator, \mathbf{G} is the additive genetic variance and covariance matrix and $\bar{w}(\bar{z}_1, \bar{z}_2)$ is the mean of individual fitness taken over the phenotype distribution of the population (Lande, 1982; Lande and Arnold, 1983; Caswell, 2006). The population responds to selection by moving uphill in the steepest direction that the selection gradient points at, $\ln \bar{w}(\bar{z}_1, \bar{z}_2)$, which is a vector of directional selection pressures (Lande, 1982).

Main results and discussion

Zero-inflated observations

In our analyzed great tit dataset, the proportion of zero observations in the number of chicks, fledglings, and recruits is 6.56%, 15.5%, 74.91% respectively. A zero-inflated Poisson model is used by Chevin et al. (2015) to estimate selection for the same population and the number of fledglings is taken as a fitness component. In the study, the zero-inflation probability is treated as a parameter instead of a separate selective episode. From a biological viewpoint, it is reasonable to assume that the complete brood failure is going through a selective process different from the expected number of fledglings. Our statistical results in paper I and III also indicate that the model where the zero-inflated observations were regarded as a separate selective episode acting on laying dates report much better model fit than the models where zero-inflated probability is taken as a model parameter.

Directional selection via complete brood failure

In paper I, the number of fledglings was partitioned into two fitness components, namely, the expected number of fledglings and the brood failure probability. The expected number of fledglings can be recognized as a straightforward extension of the conceptualization of propensity fitness, which is measured as expected rather than actual numbers of offspring (Brandon, 1978; Mills and Beatty, 1979). The best model suggested directional selection through complete brood failure and stabilizing selection via the expected number of fledglings. The direction and strength of selection via complete brood failure fluctuated over the course of study, but in most of the study years (78%) the selection favors early broods implying that females that bred late relative to the food peak were more likely to fail to raise any fledglings. Similarly, the offspring viability at each reproductive stage, from egg to chick, chick to fledgling, fledgling to recruit was split into offspring survival probability and the brood failure probability in paper III. The complete brood failure was assumed to go through directional selection in the study. The results show that selection favors early broods from stage egg to chick and implies again that early broods suffered lower probability of complete brood failure, while laying dates show no effect on complete brood failure probability from neither chick to fledgling nor from fledgling to recruit. Altogether, even though there is a much higher proportion of zero observations in number of recruit than the other episodes, directional selection operates on laying dates through complete brood failure only in the early stage of a brood, from egg to fledgling.

Stabilizing selection via expected number of fledgling (and offspring survival)

The offspring mortality is the result of malnutrition due to the mismatch between the rearing and the abundance of caterpillar peak, the main food of great tit chicks (Visser et al., 1998). Therefore, in theory, the broods laid either too early or too late relative to the peak of food resource would suffer high offspring mortality, which leads to stabilizing selection favoring the laying dates that can synchronize the chicks rearing with a narrow window of food peak. Indeed, the analyzed great tit data set supports the best model in paper I with stabilizing selection against

the model with directional selection through the expected number of fledglings on laying dates. It is thus reasonable in paper III to assume that the offspring mortality in consecutive reproductive cycle segments from egg to chick, chick to fledgling and fledgling to recruit all experienced stabilizing selection on laying dates. The properties in stabilizing fitness function (the height, location, and width) turn out to fluctuate over the course of study. The episode from chick to fledgling experienced the strongest selection implied by the smallest estimate of the width of the fitness function, compared to the other two selective episodes. Even though these three properties are assumed to be a VAR(1) process, it turns out that only the optimal laying date and width of fitness function are temporally auto-correlated and no significant cross-correlation between the fitness properties are found. It is thus safe to conclude that the annual optimal laying date and width of fitness function follow an AR(1) process respectively. The auto-correlation of optimal laying dates is estimated to as large as 0.49 and for the width of the fitness function, it is 0.64. Even though with such strong auto-correlation estimated, the simulation studies in paper I and IV suggest that the auto-correlation is probably underestimated. The temporal variation in the optimal laying dates for the different selective episodes from egg to recruit is estimated to be the same, while the variation in the width of fitness function from egg to chick is almost four times larger than the other selective episodes. In addition, the episode from fledgling to recruit estimates a much early mean optimal laying date (18.7 ± 3.1) compared to episode from egg to chick (40.7 ± 2.6) and chick to fledgling (33.5 ± 2.4). The annual overall optimal laying date calculated by maximizing the multiplication of the fitness (only for offspring survival) for the three episodes shows a close track with the optimal laying date for the third episode, from fledgling to recruit. The offspring viability (multiplication of offspring survival and non-zero inflation probability) from fledgling to recruit is also the determinant of recruit value for a specific brood and dominating the other episodes for annual reproductive success contribution. All of these imply that the cue used for timing of breeding is only available in the early breeding season, this might result from that climate change is not at constant pace through the entire breeding season, or other factors than climate have larger effects on the population outside the breeding season.

Ecological drivers of selection

One of the study aims of paper III is to identify causal mechanisms underlying temporal variation in the strength and direction of phenotypic selection on laying dates and compare the effect sizes of selective agents between the life cycle segments. We found no correlational selection on laying date and clutch size. Clutch size and laying date are negatively correlated but the correlation is weak. We found neither adult survival cost to lay broods early nor to lay big broods. The beech crop index (BCI) have lager effects on offspring survival from fledgling to recruit than from egg to fledgling, where BCI shows almost no effect. Higher BCI level is found to be positively correlated with higher female survival. Bigger clutch size is associated with higher offspring survival from egg to chick, while negatively affects offspring survival from chick to recruit. The size of the effect reduced along with the life cycle segments from egg to recruit. We also found that bigger clutches suffered a lower probability of complete brood loss from egg to chick and chick to fledgling, the effect is much stronger for the former. Not surprisingly, clutch size is negatively correlated with population density. Higher population density is found

also linked to higher offspring survival from chick to fledgling. As expected, higher population density is linked to earlier optimal laying date for offspring survival, the effect is especially strong for the episode from fledgling to recruit. The food resource peak is positively correlated with optimal laying date for each episode. In an average environment and year, the stabilizing selection strength is strongest for the episode from chick to fledgling. Higher spring temperate is associated with the wider fitness function, which suggests a weaker strength of selection. Early laying date is also linked to lower risk of complete brood loss from egg to chick but early caterpillar peak date is linked to a higher risk of complete brood loss from chick to fledgling. Breeding females differ to each in the clutch size they lay, also in the ability to survive, the ability to rear offspring successfully, and the ability to protect their broods against complete loss from egg to chick. The difference is relatively more significant for the episode from chick to fledgling.

Decreasing probability of double-brooding

The double-brooding behavior reported in our study population has been less common over the study years and the probability that a female breeds twice in a breeding season is related to the timing of her first clutch relative to the peak in caterpillar abundance (Husby et al., 2009). Indeed, we estimated the phenotypic correlation between the breeding time of first brood and liability of producing a second brood to be -0.302. Using a genetic model with parameter values estimated from the study population and a large cost of double-brooding, we show that the adaptive topography of mean population fitness exhibits two peaks at a location where there is no double-brooding or there is 100% double-brooding and the observed mean reproductive traits are overall moving towards the adaptive peak where there is no double-brooding. As long as there is no strong negative genetic correlation between the breeding time of first brood and liability of producing a second brood, the genetic model provides another possible explanation for the observed decline in the frequency of double brooding in this population in addition to the empirical study.

Laplace approximation in `tmbstan`

When using R package `tmbstan` for Bayesian inference, the built-in feature Laplace approximation to the marginal likelihood with random effects integrated out can be switched on and off. Both the simulation results and case study result in paper IV show that the Laplace approximation is accurate. In addition, turning on Laplace approximation in `tmbstan` would probably lower the computational efficiency. I conclude that only when there is a good amount of data, both `tmbstan` with and without Laplace approximation are worth trying since in this case, Laplace approximation is more likely to be accurate and may also lead to slightly higher computational efficiency. The transition parameters and scale parameters in a VAR(1) process are hard to be estimated accurately and increasing the sample size at each time point does not help in estimation, only more time points in the data contain more information on these parameters and make the likelihood dominate the posterior likelihood, thus lead to accurate estimates for them.

Conclusions and perspectives

In this thesis, we have built a statistical framework to measure fluctuating and potentially temporally auto-correlated selection, extended the framework to include more life cycle segments while taking selective forces of variation in selection into account. A simpler statistical model for estimating fluctuating selection has also been implemented in the Bayesian framework and by which we conducted simulation studies to evaluate the performance of Laplace approximation, one core feature of the Bayesian inference tool `tmbstan`. We also developed quantitative genetic models to provide a possible explanation for observed decreasing double-brooding frequency in the study population.

Either from a biological point of view or the result of statistical analysis, we found that offspring viability in the Dutch great tit population is ongoing two separate selective processes, both of which produce zero chicks/fledglings/recruits for a given brood. The nest failure experienced temporally varied directional selection and the selection generally favors early broods. The expected number of fledglings, as well as offspring survival given that the brood is successful, experienced stabilizing selection. The maximum value, optimal laying date, and width of the fitness function tend to fluctuate and auto-correlate temporally. Mother survival cost of laying eggs early is not detected. Clutch size increased along with a shift towards earlier laying date, but the effect is too small (one day earlier the laying date is, 0.0635 bigger the clutch would be) to produce a noticeable increase in clutch size even though the mean laying date has advanced around 19 days in past 50 years. We find no evidence of correlational selection on laying date and clutch size. The ecological variables, including beech crop index (BCI), population density, food peak date tend to affect one selective episode and another, in different sizes and directions. The recruit probability is the determinant of recruit value and reproductive success. The seasonal reproductive success contributed by second broods is diminishing when the first brood is laid too late provided there is no strong negative genetic correlation between the laying date of first brood and liability of attempt second brood, which provides a possible explanation for the observed decreasing frequency of double-brooding. In the state-space model, the parameters in the transition matrix and variance-covariance matrix of unobserved states are of our main interest, which are also the most difficult parameters to estimate. The simulation study in the Bayesian analysis shows that to estimate these parameters accurately, it is necessary to increase the time points in the data instead of the sample size at each time point. Laplace approximation would probably slow down the computational efficiency of MCMC especially when there is a small sample size in the data. The rule of thumb might be using Laplace approximation when you have more than 50 time points in the data.

Thanks to the new model-fitting techniques `TMB` and `tmbstan`, using state-space models to estimate a large number of parameters and random effects in complicated biological processes or ecological systems become possible even in cases where the state-space equations are highly nonlinear or non-Gaussian. By treating the phenotypical selection process as a time series and allowing a flexible covariance structure for the Gaussian fitness parameters, our method is capable of modeling different forms of variation and autocorrelation in phenotypic selection. Besides $\text{VAR}(1)$, it can also accommodate other autoregressive structures, such as $\text{VAR}(p)$ (p -order vector autoregressive process) and vector $\text{ARMA}(p,q)$ processes. Within species, there is

substantial geographic variation in the response to climate change, therefore, another direction of extending our studies could be estimating the temporal-spatial variation and correlation in fluctuating selection and investigating the causes of geographic variation in selection within species to get a better understanding of avian responses at a broader geographic scale.

Although our studies have developed applicable statistical tools for the measurement of natural selection on reproductive traits (breeding time and clutch size) through life cycle segments, however, the relationship between the timing of breeding and breeding performance is still unclear. In our studies, the clutch size has no noticeable increase in the population with a temporal shift towards earlier egg-laying. We also found no evidence of adult survival cost being laying early and selection through complete brood failure favors early broods. The probability of initializing second brood is also decreasing with delayed first brood. Take all these together, there seems no reason not to advance laying date of first brood to match the seasonal breeding time with food abundance, which is not happening in reality. One explanation could be that the timing of laying is adapted to other factors besides the timing of food supply for the chicks, or the birds are just not capable enough to track the cues of climate change. Another missing piece in our analysis is the social interaction between the phenotype (the laying date) of breeding females and males and the phenotypes of the species they associate with. At last, developing a mechanistic and theoretical understanding of the relationship between reproductive decisions and breeding performance as well as the physiological basis for these relationships are beyond the scope of our studies, but should be top priorities in extended studies since they are essential for linking the responses of birds to climate models and predicting long-term change in populations.

Bibliography

- Albertsen, C. M., Nielsen, A., and Thygesen, U. H. (2016). Choosing the observational likelihood in state-space stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(5):779–789.
- Albertsen, C. M., Whoriskey, K., Yurkowski, D., Nielsen, A., and Flemming, J. M. (2015). Fast fitting of non-gaussian state-space models to animal movement data via template model builder. *Ecology*, 96(10):2598–2604.
- Arnold, S. J. and Wade, M. J. (1984). On the measurement of natural and sexual selection: theory. *Evolution*, 38(4):709–719.
- Auger-Méthé, M., Albertsen, C. M., Jonsen, I. D., Derocher, A. E., Lidgard, D. C., Studholme, K. R., Bowen, W. D., Crossin, G. T., and Flemming, J. M. (2017). Spatiotemporal modelling of marine movement data using Template Model Builder (TMB). *Marine Ecology Progress Series*, 565:237–249.
- Barba, E., Gil-Delgado, J. A., and Monros, J. S. (1995). The costs of being late: consequences of delaying great tit parus major first clutches. *Journal of Animal Ecology*, pages 642–651.
- Bell, G. (2010). Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):87–97.
- Berg, C. W. and Nielsen, A. (2016). Accounting for correlated observations in an age-based state-space stock assessment model. *ICES Journal of Marine Science*, 73(7):1788–1797.
- Brandon, R. N. (1978). *Adaptation and evolutionary theory*. Wiley-Blackwell.
- Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, 9(2):378–400.
- Bürger, R. and Gimelfarb, A. (2002). Fluctuating environments and the role of mutation in maintaining quantitative genetic variation. *Genetics Research*, 80(1):31–46.
- Bürkner, P.-C. et al. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.
- Cadigan, N. G. (2015). A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(2):296–308.
- Calsbeek, B. (2011). Exploring variation in fitness surfaces over time or space. *Evolution*, 66(4):1126–1137.
- Cao, Y., Visser, M. E., and Tufto, J. (2019). A time-series model for estimating temporal variation in phenotypic selection on laying dates in a dutch great tit population. *Methods in Ecology and Evolution*, 10(9):1401–1411.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).
- Caswell, H. (2006). Matrix population models. *Encyclopedia of Environmetrics*, 3.

- Charlesworth, B. (1971). Selection in density-regulated populations. *Ecology*, 52(3):469–474.
- Charlesworth, B. (1993). The evolution of sex and recombination in a varying environment. *Journal of Heredity*, 84(5):345–350.
- Charnov, E. L. (1993). *Life history invariants: some explorations of symmetry in evolutionary ecology*, volume 6. Oxford University Press, USA.
- Chevin, L.-M. (2013). Genetic constraints on adaptation to a changing environment. *Evolution: International Journal of Organic Evolution*, 67(3):708–721.
- Chevin, L.-M. and Haller, B. C. (2014). The temporal distribution of directional gradients under selection for an optimum. *Evolution*, 68(12):3381–3394.
- Chevin, L.-M., Visser, M. E., and Tufto, J. (2015). Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. *Evolution*, 69(9):2319–2332.
- Clutton-Brock, T. H. (1988). *Reproductive success: studies of individual variation in contrasting breeding systems*. University of Chicago Press.
- Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.
- Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology letters*, 7(6):509–520.
- Endler, J. A. (1986). *Natural selection in the wild*. Princeton University Press.
- Engen, S., Lande, R., and Sæther, B.-E. (2011). Evolution of a plastic quantitative trait in an age-structured population in a fluctuating environment. *Evolution: International Journal of Organic Evolution*, 65(10):2893–2906.
- Evans Ogden, L. J. and Stutchbury, B. J. (1996). Constraints on double brooding in a neotropical migrant, the hooded warbler. *The Condor*, 98(4):736–744.
- Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249.
- Gamelon, M., Tufto, J., Nilsson, A. L. K., Jerstad, K., Røstad, O. W., Stenseth, N. C., and Sæther, B.-E. (2018). Environmental drivers of varying selective optima in a small passerine: A multivariate, multiepisodic approach. *Evolution*.
- Gardner, A. and Grafen, A. (2009). Capturing the superorganism: a formal theory of group adaptation. *Journal of evolutionary biology*, 22(4):659–671.
- Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.
- Geupel, G. R. and DeSante, D. F. (1990). Incidence and determinants of double brooding in wrentits. *The Condor*, 92(1):67–75.
- Hadfield, J. D. et al. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.
- Hansen, T. F. (2017). On the definition and measurement of fitness in finite populations. *Journal of theoretical biology*, 419:36–43.
- Harvey, P. H., Greenwood, P. J., and Perrins, C. M. (1979). Breeding area fidelity of great tits

- (*parus major*). *The Journal of Animal Ecology*, 48(1):305.
- Hoekstra, H. E., Hoekstra, J. M., Berrigan, D., Vignieri, S. N., Hoang, A., Hill, C. E., Beerli, P., and Kingsolver, J. G. (2001). Strength and tempo of directional selection in the wild. *Proceedings of the National Academy of Sciences*, 98(16):9157–9160.
- Hu, T., Gallins, P., and Zhou, Y.-H. (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat*, 7(1):e185.
- Husby, A., Kruuk, L. E., and Visser, M. E. (2009). Decline in the frequency and benefits of multiple brooding in great tits as a consequence of a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, 276(1663):1845–1854.
- Jacobs, A. C., Reader, L. L., and Fair, J. M. (2013). What determines the rates of double brooding in the western bluebird? ¿ qué determina las tasas de nidada doble en sialia mexicana? *The Condor*, 115(2):386–393.
- Johnson, D. S., London, J. M., Lea, M.-A., and Durban, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology*, 89(5):1208–1215.
- Jonsen, I. D., Flemming, J. M., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology*, 86(11):2874–2880.
- Kingsolver, J. G., Hoekstra, H. E., Hoekstra, J. M., Berrigan, D., Vignieri, S. N., Hill, C., Hoang, A., Gibert, P., and Beerli, P. (2001). The strength of phenotypic selection in natural populations. *The American Naturalist*, 157(3):245–261.
- Kristensen, K. (2018). *MCMC Sampling from 'TMB' Model Object using 'Stan'*. R package version 1.0.1.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J., and Bell, B. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.
- Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution*, 33(1Part2):402–416.
- Lande, R. (1982). A quantitative genetic theory of life history evolution. *Ecology*, 63(3):607–615.
- Lande, R. and Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, 37(6):1210–1226.
- Lande, R. and Shannon, S. (1996). The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution*, 50(1):434–437.
- Liker, A., Freckleton, R. P., Remeš, V., and Székely, T. (2015). Sex differences in parental care: Gametic investment, sexual selection, and social environment. *Evolution*, 69(11):2862–2875.
- Lindén, M. (1988). Reproductive trade-off between first and second clutches in the great tit *parus major*: an experimental study. *Oikos*, pages 285–290.
- Linnen, C. R. and Hoekstra, H. E. (2009). Measuring natural selection on genotypes and phenotypes in the wild. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 74, pages 155–168. Cold Spring Harbor Laboratory Press.
- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.
- Magnusson, A., Skaug, H., Nielsen, A., Berg, C., Kristensen, K., Maechler, M., van Bentham, K., Bolker, B., Brooks, M., and Brooks, M. M. (2017). Package ‘glmmTMB’. *R Package Version 0.2. 0*.
- Martin, T. E. and Briskie, J. V. (2009). Predation on dependent offspring: a review of the consequences for mean expression and phenotypic plasticity in avian life history traits. *Annals of the New York Academy of Sciences*, 1168(1):201–217.
- McClerey, R., Clobert, J., Julliard, R., and Perrins, C. (1996). Nest predation and delayed cost of reproduction in the great tit. *Journal of Animal Ecology*, pages 96–104.
- McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., and Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs*, 82(3):335–349.
- Merilä, J., Sheldon, B., and Kruuk, L. (2001). Explaining stasis: microevolutionary studies in natural populations. *Genetica*, 112(1):199–222.
- Mills, S. K. and Beatty, J. H. (1979). The propensity interpretation of fitness. *Philosophy of Science*, 46(2):263–286.
- Monnahan, C. C. and Kristensen, K. (2018). No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PloS one*, 13(5):e0197954.
- Monnahan, C. C., Thorson, J. T., and Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3):339–348.
- Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, 85(9):2436–2445.
- Morrissey, M. B. and Hadfield, J. D. (2012). Directional selection in temporally replicated studies is remarkably consistent. *Evolution: International Journal of Organic Evolution*, 66(2):435–442.
- Mulvihill, R. S., Latta, S. C., and Newell, F. L. (2009). Temporal constraints on the incidence of double brooding in the louisiana waterthrush. *The Condor*, 111(2):341–348.
- Nagy, L. R. and Holmes, R. T. (2004). Factors influencing fecundity in migratory songbirds: is nest predation the most important? *Journal of Avian Biology*, 35(6):487–491.
- Newton, I. et al. (1989). *Lifetime reproduction in birds*. Academic Press.
- Parejo, D. and Danchin, E. (2006). Brood size manipulation affects frequency of second clutches in the blue tit. *Behavioral ecology and sociobiology*, 60(2):184–194.
- Pedersen, M. W., Berg, C. W., Thygesen, U. H., Nielsen, A., and Madsen, H. (2011). Estimation methods for nonlinear state-space models in ecology. *Ecological Modelling*, 222(8):1394–1400.
- Perrins, C. M. and McCleery, R. H. (1989). Laying dates and clutch size in the great tit. *The Wilson Bulletin*, pages 236–253.

- Price, G. R. (1970). Selection and covariance. *Nature*, 227(5257):520–521.
- Price, T. and Liou, L. (1989). Selection on clutch size in birds. *The American Naturalist*, 134(6):950–959.
- Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of computational and Graphical Statistics*, 9(1):141–157.
- Reed, T. E., Gienapp, P., and Visser, M. E. (2016). Testing for biases in selection on avian reproductive traits and partitioning direct and indirect selection using quantitative genetic models. *Evolution*, 70(10):2211–2225.
- Reed, T. E., Grøtan, V., Jenouvrier, S., Sæther, B.-E., and Visser, M. E. (2013a). Population growth in a wild bird is buffered against phenological mismatch. *Science*, 340(6131):488–491.
- Reed, T. E., Jenouvrier, S., and Visser, M. E. (2013b). Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology*, 82(1):131–144.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.
- Saccheri, I. and Hanski, I. (2006). Natural selection and population dynamics. *Trends in Ecology & Evolution*, 21(6):341–347.
- Sæther, B.-E. and Bakke, Ø. (2000). Avian life history variation and contribution of demographic traits to the population growth rate. *Ecology*, 81(3):642–653.
- Sæther, B.-E., Visser, M. E., Grøtan, V., and Engen, S. (2016). Evidence for r-and k-selection in a wild bird population: a reciprocal link between ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 283(1829):20152411.
- Santema, P. and Kempenaers, B. (2018). Complete brood failure in an altricial bird is almost always associated with the sudden and permanent disappearance of a parent. *Journal of Animal Ecology*, 87(5):1239–1250.
- Schlüter, D. (1988). Estimating the form of natural selection on a quantitative trait. *Evolution*, 42(5):849–861.
- Schlüter, D., Price, T. D., and Rowe, L. (1991). Conflicting selection pressures and life history trade-offs. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 246(1315):11–17.
- Siepielski, A. M., DiBattista, J. D., and Carlson, S. M. (2009). It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecology letters*, 12(11):1261–1276.
- Siepielski, A. M., DiBattista, J. D., Evans, J. A., and Carlson, S. M. (2010). Differences in the temporal dynamics of phenotypic selection among fitness components in the wild. *Proceedings of the Royal Society B: Biological Sciences*, 278(1711):1572–1580.
- Siiämäki, P. (1998). Limitation of reproductive success by food availability and breeding time in pied flycatchers. *Ecology*, 79(5):1789–1796.
- Skaug, H., Fournier, D., and Nielsen, A. (2013). glmmADMB: generalized linear mixed models

using AD Model Builder–R Package.

- Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMMSS in R*. Chapman and Hall/CRC.
- Stearns, S. C. (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften*, 87(11):476–486.
- Stinchcombe, J. R., Kelley, J. L., Conner, J. K., and Freckleton, R. (2017). How to measure natural selection. *Methods in Ecology and Evolution*, 8(6).
- Team, S. D. (2017). Stan modeling language: User’s guide and reference manual. *Version 2.17.0*.
- Thomson, C. E. and Hadfield, J. D. (2017). Measuring selection when parents and offspring interact. *Methods in Ecology and Evolution*, 8(6):678–687.
- Townsend, A. K., Sillett, T. S., Lany, N. K., Kaiser, S. A., Rodenhouse, N. L., Webster, M. S., and Holmes, R. T. (2013). Warm springs, early lay dates, and double brooding in a north american migratory songbird, the black-throated blue warbler. *PLoS One*, 8(4):e59467.
- Tufto, J. (2015). Genetic evolution, plasticity, and bet-hedging as adaptive responses to temporally autocorrelated fluctuating selection: a quantitative genetic model. *Evolution*, 69(8):2034–2049.
- Verboven, N., Tinbergen, J. M., and Verhulst, S. (2001). Food, reproductive success and multiple breeding in the great tit parus major. *Ardea*, 89(2):387–406.
- Verboven, N. and Verhulst, S. (1996). Seasonal variation in the incidence of double broods: the date hypothesis fits better than the quality hypothesis. *Journal of Animal Ecology*, pages 264–273.
- Verboven, N. and Visser, M. E. (1998). Seasonal variation in local recruitment of great tits: the importance of being early. *Oikos*, pages 511–524.
- Visser, M., Van Noordwijk, A., Tinbergen, J., and Lessells, C. (1998). Warmer springs lead to mistimed reproduction in great tits (parus major). *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1408):1867–1870.
- Wang, G. (2007). On the latent state estimation of nonlinear population dynamics using Bayesian and non-Bayesian state-space models. *Ecological Modelling*, 200(3):521–528.
- Wei, W. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. Pearson Addison Wesley, 2nd edition.
- Wesołowski, T. (2002). Anti-predator adaptations in nesting marsh tits parus palustris: the role of nest-site security. *Ibis*, 144(4):593–601.
- Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.
- Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, 27(8):1–25.
- Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

A time-series model for estimating temporal variation in phenotypic selection on laying dates in a Dutch great tit population

Yihan Cao¹ | Marcel E. Visser² | Jarle Tufto¹

¹Centre for Biodiversity Dynamics,
Department of Mathematical
Sciences, Norwegian University of Science
and Technology, Trondheim, Norway

²Department of Animal Ecology,
Netherlands Institute of Ecology
(NIOO-KNAW), Wageningen, Netherlands

Correspondence
Yihan Cao
Email: yihan.cao@ntnu.no

Handling Editor: Thomas Hansen

Abstract

1. Temporal and spatial variation in phenotypic selection due to changing environmental conditions is of great interest to evolutionary biologists, but few existing methods estimating its magnitude take into account the temporal autocorrelation.
2. We use state-space models (SSMs) to analyse phenotypic selection processes that cannot be observed directly and use Template Model builder (TMB), an R package for computing and maximizing the Laplace approximation of the marginal likelihood for SSM and other complex, nonlinear latent variable model via automatic differentiation. Using a long-term great tit (*Parus major*) dataset, we fit several SSMs and conduct model selection based on Akaike information criterion (AIC) to assess the support for fluctuated directional or autocorrelated stabilizing selection on breeding time of the great tit population.
3. Our results show that there is directional selection on the probability of breeding failure, and stabilizing selection on the mean number of fledglings. This selection for early laying date is consistent with a previous study of the same population. We also estimate the variation and autocorrelation in other parameters of the fitness functions, including the width and height, and found the height and location of annual fitness function are autocorrelated with significant variation, while the width can be assumed to constant over time.
4. Using TMB to fit SSMs, we are able to estimate additional parameters compared to previous methods, all without requiring a substantial increase in computational resources. Furthermore, our specification of complex nonlinear model structure benefits greatly from the flexibility of model formulation with TMB. Therefore, our approach could be directly applied to estimating even more complicated phenotypic selection processes induced by environmental change for other species.

KEY WORDS

fluctuating selection, Gaussian fitness function, generalized linear mixed model, state-space model, template model builder, zero-inflated Poisson regression

1 | INTRODUCTION

Fluctuating selection resulting from environmental variation has been of long-lasting interest. Empirical and theoretical research have documented that natural populations respond to varying selection through various mechanism including conventional Darwinian genetic evolution (Lande & Shannon, 1996), evolution of phenotypic plasticity (Scheiner, 1993; Van Tienderen & Koelewijn, 1994), evolution of genetic polymorphism (reviewed by Hedrick, 2006; Bell, 2010), genetic variance (Barton & Keightley, 2002), evolution of the phenotypic variance (Zhang & Hill, 2005) including diversifying bet-hedging (Bull, 1987; Cohen, 1966; Svardal, Rueffler & Hermission, 2011) or combinations of these response modes (Tufto, 2015). Importantly, the relative magnitudes of these different responses depend on the temporal autocorrelation in selective optima. Even though the phenotypic traits typically evolve through natural selection to match the environmental conditions to maximize fitness (Futuyma, 2006), phenotypic adaptation through genetic evolution is limited by the amount of genetic variance in the trait under selection, which might lead to mistiming between the mean phenotype and the phenotypic optimum (Lande & Shannon, 1996). Adaptive tracking through phenotypic plasticity acting in conjunction with genetic evolution may also be limited by factors such as imperfect cue reliability (Post & Forchhammer, 2007; Gienapp, Reed & Visser, 2014) or parental energetic costs (Visser, Marvelde & Lof, 2012; Visser et al., 2015).

There are few studies estimating the temporal variability and autocorrelation of phenotypic selection in spite of the importance. The variance in phenotypic selection in previous studies was usually estimated by computing the variance of the strength of selection using selection gradients estimated separately at each time point (reviewed by Siepielski, DiBattista & Carlson, 2009), which reflects both sampling error and real variation in selection (Morrissey & Hadfield, 2012). Among the previous empirical studies accounting for the sampling error of variation, Calsbeek (2011) presented a nonparametric analysis in exploring the variation of fitness surfaces over time or space, but such nonparametric estimates are difficult to relate to parameters appearing in theoretical models. In contrast, using a log-quadratic generalized linear mixed model (GLMM) with a random effect on the regression slope implemented using integrated nested Laplace approximations (INLA) (Rue, Martino & Chopin, 2009), Chevin, Visser and Tufto (2015) estimated yearly fluctuations and autocorrelation in optima of a Gaussian fitness function. However, INLA and GLMMs in general are restricted to cases where the predictor is linear in parameters and random effects. Using instead the more flexible framework of Template Model Builder (TMB) (Kristensen, Nielsen, Berg, Skaug & Bell, 2015), Gamelon et al. (2018) fitted a model of fluctuating selection via several non-overlapping selection episodes with nonlinear random effects added directly on the location of the fitness optima and on the peak of the fitness function. This model form is not feasible within the framework of INLA or GLMMs (see Gamelon et al., 2018, Appendix A for a technical discussion).

Here, we extend the approach taken in Chevin et al. (2015) and Gamelon et al. (2018) in several new ways. First, instead of assuming a fixed zero-inflation parameter for modelling the number of fledglings as in Chevin et al. (2015), we model the zero-inflation probability using a separate linear (or nonlinear) predictor. This leads to a model with selection via zero-inflation and via the Poisson mean, although occurring during the same interval. As with multi-episodic selection more generally (Gamelon et al., 2018), selection through two episodes can involve the same or different biological processes. Second, in addition to random effects on the peak and location of the fitness optimum as in Gamelon et al. (2018), we also allow the width of the fitness function to vary between years, with all three properties of the Gaussian fitness function jointly following a vector autoregressive process. Such variation in the width is of theoretical importance for the evolution of the phenotypic variance (Zhang & Hill, 2005) and for the evolutionary stability of the additive genetic variance-covariance matrix (Revell, 2007). Third, instead of treating the total number of fledglings from all broods laid by a female in a particular year as the sample unit and estimating stabilizing selection on onset of breeding via its effect on the sum of number of fledglings from all broods as in Chevin et al. (2015), we treat the number of fledglings from each brood as the sample unit and fit the model under the assumption that the expected number of fledglings depend on the laying date according to the same Gaussian fitness function for all broods. In addition to increased statistical power, this has the advantage that the parameters relate directly to theoretical models for the joint evolution of multiple brooding and onset of breeding (Tufto, Cao and Visser, submitted manuscript). Fourth, as an alternative to stabilizing selection, we allow each episode (here selection via zero-inflation and via the Poisson mean) to instead involve directional selection. As in Gamelon et al. (2018), we implement our method using TMB (Kristensen et al., 2015), an R package providing a comprehensive framework for fast fitting nonlinear, complex, latent variable models.

2 | MATERIALS AND METHODS

2.1 | Study population

The data analysed come from a natural population of great tits (*Parus major*) at the Hoge Veluwe National Park in the Netherlands ($52^{\circ}02' - 52^{\circ}07'N$, $5^{\circ}51' - 5^{\circ}32'E$). Female great tits usually start reproduction in the second calendar year of life (Perrins, 1979) and are capable of producing a second and very rarely, a third brood in a season. The analysed dataset consists of 5892 records of 3257 females breeding in 61 years from 1955 to 2015. Unlike the previous studies on the same population (e.g. Reed, Jenouvrier & Visser, 2013), we kept the data from the 1991 breeding season when a late frost led to a very late caterpillar food peak (Visser, Noordwijk, Tinbergen, & Lessells 1998) and we expected a very late optimum estimate for this breeding season. Laying dates are presented as the number of days after March 31 (day 1 = April 1, day 31 = May 1). The number of fledglings for each visited brood

was counted and the mother of each brood was identified (3257 breeding mothers in our analysed data). The average number of breeding records per known female was 1.81. See Supporting Information for more details on study population and fieldwork procedures.

2.2 | Model formulation

We formulated a statistical model that takes into account temporally fluctuated stabilizing selection and used laying date as the focal trait that selection operates on. We also considered alternative models assuming fluctuated directional selection. We take the number of individuals surviving to fledglings as the measure of fecundity component of fitness and it is assumed to follow a zero-inflated Poisson distribution instead of a Poisson distribution due to the high probability of clutch failure (around 15.7% in our analysed dataset, clutch failure in this study means that no single chick survived to fledgling). In addition, previous studies showed (e.g. Reed et al., 2013; Townsend et al., 2013) that the relative contribution to fitness from each brood, at individual level, is determined by the food abundance at the time each brood is raised. We therefore assume that the expected number of fledglings and the probability of clutch failure potentially depends on laying dates in the same way for first, replacement (first broods failed) and second (first broods succeeded) broods via the same Gaussian fitness function. We present our approach using selection on the number of fledglings, but it can be applied to any selection episodes, such as viability, fertility selection or overall selection through lifetime fitness for species with non-overlapping generations.

We assume that the number of fledglings Y_i ($i = 1, 2, \dots, n$) from the i th brood follow a zero-inflated Poisson (ZIP) distribution. Such random variables can be represented as a product $Y_i = I_i X_i$ where

$$\begin{aligned} I_i | p_i &\sim \text{Bernoulli}(1 - p_i), \\ X_i | I_i = 1, w_i &\sim \text{Poisson}(w_i). \end{aligned} \quad (1)$$

Here, p_i is the probability of zero-inflation (complete brood failure), w_i is the Poisson mean and i is the index for all of the broods in our analysed dataset, $i = 1, 2, \dots, 5892$. Using the law of total expectation, the overall fitness contribution from brood i is then

$$\begin{aligned} E(Y_i | p_i, w_i) &= E(X_i I_i | p_i, w_i) \\ &= E(X_i I_i | p_i, w_i, I_i = 1) P(I_i = 1) + E(X_i I_i | p_i, w_i, I_i = 0) P(I_i = 0) \\ &= E(X_i | I_i = 1, w_i) E(I_i | p_i). \end{aligned} \quad (2)$$

The decomposition of the left-hand side into the two factors on the right-hand side shows that the zero-inflation part I_i can be interpreted as a separate selection episode, which we refer to as episode P for short in this study. Similarly, the Poisson part X_i is referred to as episode W in the rest of this paper.

We consider two selection modes: fluctuating stabilizing selection and fluctuating directional selection. In the fluctuating stabilizing selection model, the zero-inflation probability p_i and the Poisson mean w_i are determined by the same process, driven by deviation from the optimal onset of breeding. In addition, we assume that p_i

is linked to covariates of interest via a logit link function while for w_i via a log link function. Therefore, $\text{logit}(1 - p_i)$ and $\ln w_i$ are given by models of the same form:

$$\text{logit}(1 - p_i) = \eta_{p,t}^{(\alpha)} - \frac{(z_i - \eta_{p,t}^{(\theta)})^2}{2(e^{\eta_{p,t}^{(\omega)}})^2} + \tau_p^m \epsilon_{j(i)}, \quad (3)$$

and

$$\ln w_i = \eta_{w,t}^{(\alpha)} - \frac{(z_i - \eta_{w,t}^{(\theta)})^2}{2(e^{\eta_{w,t}^{(\omega)}})^2} + \tau_w^m \epsilon_{j(i)}. \quad (4)$$

Here, $\eta_{p,t}^{(\alpha)}$, $\eta_{p,t}^{(\theta)}$ and $\eta_{p,t}^{(\omega)}$, $t = 1, 2, \dots, 61$ are parameters determining maximum fitness (indicated by superscript α), optimal laying dates (indicated by θ) and widths of fitness function (indicated by ω) of brood i in year t respectively for $\text{logit}(1 - p_i)$. Similar explanations apply to the equation of $\ln w_i$. The variable z_i is the laying date of the i th brood. The term $\epsilon_{j(i)} \sim N(0, 1)$, $j = 1, 2, \dots, J$ (where J is total number of unique females) is a random effect included to model extra variation between the mothers and assumed to be same for the two episodes, but the magnitude of the effects on the two episodes are potentially different, subscript p, w thereby allow standard deviations of mother effect τ_p^m , τ_w^m to differ between episode P and W .

The maximum fitness, optimal laying date and width of fitness function in the two episodes are assumed to have a constant difference c_α , c_θ and c_ω ($\eta_{w,t}^{(\alpha)} = \eta_{p,t}^{(\alpha)} + c_\alpha$, $\eta_{w,t}^{(\theta)} = \eta_{p,t}^{(\theta)} + c_\theta$ and $\eta_{w,t}^{(\omega)} = \eta_{p,t}^{(\omega)} + c_\omega$) and we therefore model $\eta_{p,t}^{(\alpha)}$, $\eta_{p,t}^{(\theta)}$ and $\eta_{p,t}^{(\omega)}$ by the three stochastic processes

$$\begin{aligned} \eta_{s,t}^{(\alpha)} &= \bar{\alpha}_s + \alpha_t, \\ \eta_{s,t}^{(\theta)} &= \bar{\theta}_s + \theta_t, \\ \eta_{s,t}^{(\omega)} &= \bar{\omega}_s + \omega_t, \end{aligned} \quad (5)$$

where index s takes values from (P, W) indicating the two episodes respectively. Parameters $\bar{\alpha}_s$, $\bar{\theta}_s$, $\bar{\omega}_s$ are the means of the three processes. More assumptions in terms of stochastic processes α_t , θ_t , ω_t are made. They are assumed to follow a first-order vector autoregressive VAR(1) process

$$\begin{bmatrix} \alpha_t \\ \theta_t \\ \omega_t \end{bmatrix} = \Phi \begin{bmatrix} \alpha_{t-1} \\ \theta_{t-1} \\ \omega_{t-1} \end{bmatrix} + \mathbf{w}_t, \quad (6)$$

where Φ is a 3×3 matrix of autoregressive coefficients and \mathbf{w}_t is multivariate normal zero-mean white noise with variance-covariance matrix Σ . Correlation between α_t , θ_t and ω_t are determined through off-diagonal entries in both Σ and Φ . Possible model alternatives are obtained by making Φ and Σ both diagonal, such that α_t , θ_t and ω_t simplify to independent AR(1) processes. If all entries of Φ are zero, α_t , θ_t and ω_t are independent and identically distributed white noise processes. Alternatively, we model each episode as fluctuating directional selection, which can be described by a GLMM with annual random intercept and slope:

$$\begin{aligned} \text{logit}(1 - p_i) &= \beta_p^{(0)} + u_{p,t}^{(0)} + (\beta_p^{(1)} + u_{p,t}^{(1)}) z_i + \tau_p^m \epsilon_{j(i)}; \\ \ln w_i &= \beta_w^{(0)} + u_{w,t}^{(0)} + (\beta_w^{(1)} + u_{w,t}^{(1)}) z_i + \tau_w^m \epsilon_{j(i)}. \end{aligned} \quad (7)$$

In this model, $\beta_p^{(0)}$ and $\beta_p^{(1)}$ are fixed intercept and slope respectively for episode P, random intercepts and slopes are denoted by $u_{p,t}^{(0)}$ and $u_{p,t}^{(1)}$, which account for the variation among years. These random effects are assumed to be multivariate normal:

$$\begin{pmatrix} u_{p,t}^{(0)} \\ u_{p,t}^{(1)} \end{pmatrix} \sim N \left(\mathbf{0}, \begin{pmatrix} (\sigma_{0,p})^2 & \rho_p \sigma_{0,p} \sigma_{1,p} \\ \rho_p \sigma_{0,p} \sigma_{1,p} & (\sigma_{1,p})^2 \end{pmatrix} \right).$$

Similar explanation applies to the alternative model for episode W ($\ln w_i$). As before, z_i , τ_p^m , τ_w^m and $\epsilon_{j(i)}$ have same interpretations as that in Equations 3 and 4.

Since our statistical method relies on model selection, the candidate models we tested include different assumptions for α_t , θ_t and ω_t , or different selection patterns for episode P and W, among many others.

2.3 | Model selection and inference

All model alternatives were implemented using R package TMB. Briefly, based on a C++ function computing the joint density of the observed data and unobserved random effects, TMB computes the Laplace approximation of the marginal likelihood of the observed data. This is then maximized numerically to obtain maximum likelihood estimates of model parameters and approximate standard errors based on information theory.

We fitted in total 43 different alternative models. Among the candidate models, each selection episode P and W maybe equipped with either directional selection or stabilizing selection. For the directional selection mode, we tested models with only fixed effects, with random intercepts and with both random intercepts and random slopes. For the stabilizing selection mode, the fitness parameters α_t , θ_t and ω_t were either considered as constant, as three independent AR(1) processes, as jointly following a VAR(1) process, or combinations of them.

Our model selection relies on the measurement of data support for the different models which vary in degree of complexity. We use Akaike information criterion (AIC; Akaike, 1973) based on the observed Fisher information as a model selection criterion (see Burnham & Anderson, 2003 for more details about AIC). The model with lowest AIC value was selected as the best model and the estimates of all parameters together with their approximate standard errors were obtained. All the source code of this study are archived and accessible online.

3 | CASE STUDY RESULTS AND DISCUSSION

3.1 | Model selection procedure

As introduced in section 2.3, in total, 43 candidate models were tested. For brevity, only the selected model and its neighbour models are listed in Table 1. The model numbering is consistent with the model updating sequence in our R code. Updating procedure from

model 1 to model 8b can be found in Supporting Information. Based on the best model selected (model 9), the differences of AIC value for each model from the selected model are calculated and listed in column Δ AIC, along with the difference in the number of parameters (Δp). Model 9 with directional selection in episode P and stabilizing selection in episode W is the best model.

Model 10 with directional selection via both episode P and W does not improve the model fit. To guarantee that model 9 is indeed the best one among all the candidate models, model 11 to model 14i are neighbour models updated around model 9 for comparison purpose, but none of them improves the model fit. It is worth noting that the performance of model 14g with fixed ω_t is only slightly worse than our selected model, implying a constant ω_t assumption in our study would be reasonable. The estimates of parameters from the selected model (model 9) and from model 14g with constant ω_t are listed in the Supporting Information for comparison.

We also carried out a simulation study (see Supporting Information) to explore the power of our model selection technique in identifying our best model especially against model 14g and 14i. We concluded from the simulation study that our model selection technique has around 80% probability to distinguish the model with fixed ω_t from the one with random ω_t when the variation scale of ω_t being 0.2. This further implies that model 14g might be as good as our selected model. The simulation study also showed that a weak mother effect (e.g. the standard deviation of random mother effects is 0.05) is hard to detect. However, since our selected model reports 6.44 lower AIC values with the estimate of standard deviation of mother effects being 0.041 in episode W, we have confidence in the mother effects in the underlying ‘true’ model. The remaining challenge is that there is no strong evidence for model 9 outperforms model 13, we thus should be cautious when interpreting estimates of correlations between the errors for α_t , θ_t and ω_t .

The selected model (model 9) has stabilizing selection via episode W, directional selection via P with annual correlated random intercepts and slopes given by

$$\begin{aligned} \text{logit}(1-p_i) &= \beta_p^{(0)} + u_{p,t}^{(0)} + \left(\beta_p^{(1)} + u_{p,t}^{(1)} \right) z_i + \tau_p^m \epsilon_{j(i)}, \\ \ln w_i &= \eta_{w,t(i)}^{(\alpha)} - \frac{\left(z_i - \eta_{w,t(i)}^{(\alpha)} \right)^2}{2 \left(e^{u_{w,t(i)}^{(\alpha)}} \right)^2} + \tau_w^m \epsilon_{j(i)}. \end{aligned} \quad (8)$$

Furthermore, the selected model supports VAR(1) process of α_t , θ_t and ω_t in the episode W. However, the three processes are correlated through errors instead of the transition matrix Φ . Mother effects are significant in both episodes. More details about the parameter estimates are given in next section 3.2.

3.2 | Directional selection via probability of clutch failure

Our selected model (Equation 8) indicates directional selection via episode P with annual random intercept ($u_{p,t}^{(0)}$) and slope ($u_{p,t}^{(1)}$). The estimates

TABLE 1 A part of model selection procedure of phenotypic selection on breeding time of great tits. The order of models listed below is accordance with the order of models fitting, from model 9 to 14i. ΔAIC and Δp is the difference in AIC and number of parameters p between each model and the best model (model 9). The column of description gives the details of the updated model based on the previous ones. For simplification, the probability of successful brooding is denoted as episode P and the mean number of fledglings episode as W . The updating procedure from model 1 to model 8a can be found in Supporting Info

Model	ΔAIC	Δp	Description
Directional selection via episode P and stabilizing selection via episode W			
9	0	0	The selected model formulated as Equation 8
Directional selection for both episode P and W			
10	104.11	-5	The model formulated as Equation 7 with correlated random intercepts and slopes
Model 9 is the best model so far, test neighbour models with minor changes based on model 9			
11	4.075	7	Add all entries of Φ back
12	14.32	-3	Keep only significant entries in Φ and significant correlations between the errors for α_t , θ_t and ω_t
13	0.54	-2	Keep only significant correlations between the errors for α_t , θ_t and ω_t
Model 9 is still the best model so far, test models with all possible specifications for α_t , θ_t and ω_t , and remove mother effect from each episode			
14	210.47	-6	Change random α_t into fixed, θ_t and ω_t are random
14a	203.98	-4	Change random α_t into fixed, θ_t and ω_t are AR(1)
14b	96.99	-6	Change random θ_t into fixed, α_t and ω_t are random
14c	85.47	-4	Change random θ_t into fixed, α_t and ω_t are AR(1)
14d	24.84	-4	Change random ω_t into fixed, α_t and θ_t are AR(1)
14e	28.64	-2	Change random ω_t into fixed, VAR(1) α_t and θ_t
14f	7.1	-1	Change random ω_t into fixed, VAR(1) α_t and θ_t , add correlation to the errors of α_t and θ_t
14g	3.27	-3	Change random ω_t into fixed, VAR(1) α_t and θ_t with significant entries in Φ , add correlation to the errors of α_t and θ_t
14h	21.1	-1	Remove mother effect from episode P
14i	6.44	-1	Remove mother effect from episode W

TABLE 2 Estimates (standard errors) and corresponding 95% confidence intervals of model parameters from the selected model (i.e. model 9 of Table 1, only for selection via the probability of successful brooding)

Parameter	Meaning	Estimate (SE)	95% CI
$\beta_p^{(0)}$	Fixed intercept	2.946 (0.220)	(2.515, 3.377)
$\beta_p^{(1)}$	Fixed slope	-0.025 (0.005)	(-0.035, -0.015)
$\sigma_{1,p}$	SD of random slopes	0.032 (0.004)	(0.024, 0.040)
r_p^m	SD of mother effect	0.701 (0.092)	(0.520, 0.881)
ρ_p	Correlation between random intercepts and slopes	-0.827 (0.054)	(-0.933, -0.720)

of parameters of our interest are listed in Table 2. We estimated the fixed slope, the mean of the annual selection gradient to $\hat{\beta}_p^{(1)} = -0.025$ (red dashed line on the right panel of Figure 1). Given a standard deviation of the random slopes estimated to $\hat{\sigma}_{1,p} = 0.032$, corresponding to selection for early laying 78% of the time, the distribution of selection gradients is shown with the black line in the right panel of Figure 1,

which implies that over 22% of the study years experienced positive selection, therefore, favoured late broods. The left panel of Figure 1 shows the annual selection gradient together with error bars representing \pm one standard error. The selection favoured early broods in 82% (note that the 78% is obtained with selection distribution while 82% with temporal estimated selection) of the study years, as can be seen from the left panel that most of the selection gradients fall below 0.

This result agrees with the finding from Reed et al. (2013) that females that breed late relative to the food peak (influenced by temperature, see Visser, Holleman and Gienapp (2006)) were more likely to fail to raise any fledglings. Perrins (1965) states that there is a higher proportion of predation in the later part of the season and the young of the later broods are more vulnerable to the predators since the young in the later broods are more noisy and lighter. Maziarz, Wesolowski, Hebda, Cholewa and Broughton (2016) shows that nest losses are mostly due to predation (69% nest failures of a great tit population in Poland) and the risk of nest failure varied with nest cavity attributes. To explore which biotic and abiotic factors best explain the sign and variation in annual directional selection via the probability of successful brooding, more data information concerning these factors are required and this would be one among other interesting expansions of this study. In this selection episode, mother effects contribute to explaining the variation of successful-brooding probabilities and the estimate of the standard deviation r_p^m is 0.701, as shown in Table 2.

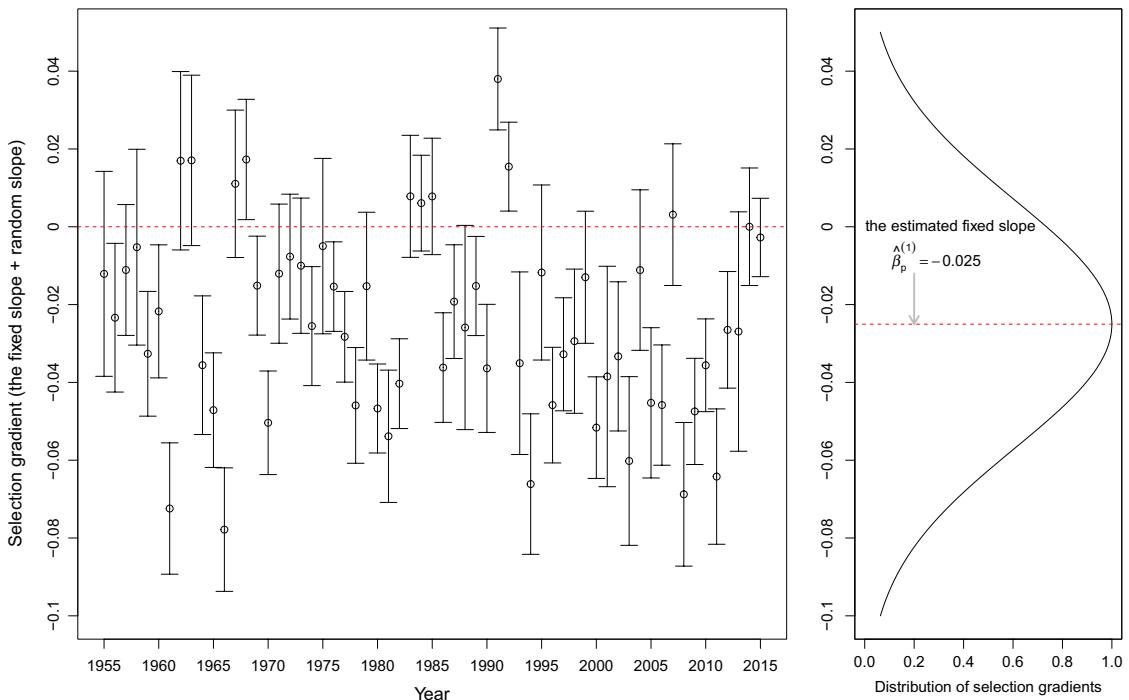


FIGURE 1 Annual directional selection gradient (left panel, defined as the sum of the fixed slope and annual random slope) associated with error bars representing one standard error and the distribution of it (right panel) for episode P (selection on laying date via the probability of successful brooding). The red dashed line on the left plot is an indication of 0 and on the right plot is estimated fixed slope (-0.025)

3.3 | Stabilizing selection via the mean number of fledglings

Our selected model indicates that stabilizing selection acts via the mean of the Poisson component. Parameter estimates of the Gaussian fitness function in Equation 4, and the estimates for the parameters involved in the VAR(1) α_t , θ_t and ω_t , along with their confidence intervals are shown in Table 3.

The estimates of the mean of maximum fitness (\bar{x}) and optimum ($\bar{\theta}$) are 2 (exponent with base e approximates to 7 fledglings) and 18.227 (approximately 18th of April) respectively. Our estimate for the width of the fitness function is much wider than that from Chevin et al. (2015) (47.395 vs. 24.11 days), in which the sum of the fledglings from multiple broods instead of the single brood was treated as the sample unit and the lay dates of only first broods (with a much narrower range) were used. We doubt that the distribution of this summation of multiple broods is well approximated by a Gaussian function and therefore we modelled the number of fledglings from each brood separately, and the second broods were laid in the late breeding season and this might be the reason of a wider fitness function being estimated with our selected model.

The estimates of the standard deviation of α_t , θ_t and ω_t are 0.176, 21.180 and 0.205 respectively. The estimate of standard deviation

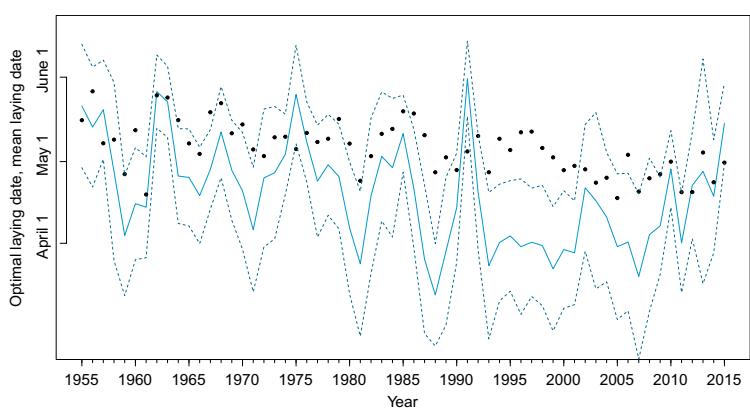
for θ_t is slightly larger than that from Chevin et al. (2015) (21.18 vs. 11.3 days) and this might partly result from the different datasets we used. In Chevin et al. (2015), the data before 1973 were excluded from their analysis and we therefore also fit the selected model with data only after 1973 for a fairer comparison. It turned out that the estimates with both full and partial datasets are quite close, while the estimates with full data generally have less uncertainty (narrower 95% confidence intervals). The detailed comparison can be found in Supporting information. The estimated variance in ω_t (0.042) translates to a coefficient of variation for e^{ω_t} of $\sqrt{e^{0.042} - 1} = 0.207$, that is, quite large fluctuation in the width of the fitness function. When conducting model selection we fitted a model with fixed ω_t (model 14g in Table 1) over study period, however, it turned out the model fit did not improve much when ω_t is taken random as in our selected model. In addition, by comparing the standard deviations of parameter estimates from the models with fixed and random ω_t reported in Supporting Information, we find that uncertainties of parameter estimates are comparable. These imply that the whole analysis would not change much if in our study the constant ω_t assumption was made.

The autocorrelation estimates of α_t ($\hat{\phi}_{\alpha,\alpha}$) and θ_t ($\hat{\phi}_{\theta,\theta}$) are 0.334 and 0.524, respectively, but ω_t is not autocorrelated in our selected model. The estimate of $\phi_{\theta,\theta}$ in Chevin et al. (2015) was 0.2472 with a wide 95% confidence interval (-0.1745, 0.626). While our selected

TABLE 3 Estimates (standard errors) and corresponding 95% confidence intervals of model parameters with the selected model (only for selection via the mean number of fledglings)

Parameter	Meaning	Estimate (SE)	95% CI
$\bar{\alpha}$	Mean of process $\eta_{w,t}^{(\alpha)}$	2.000 (0.036)	(1.929, 2.071)
$\bar{\theta}$	Mean of process $\eta_{w,t}^{(\theta)}$	18.227 (5.826)	(6.808, 29.647)
e^{ϕ} (days)	Mean of process $e^{f_{w,t}^{(\phi)}}$	47.395 (3.234)	(41.056, 53.734)
$\gamma_{\alpha,\alpha}$	SD of α_t	0.176 (0.024)	(0.129, 0.224)
$\gamma_{\theta,\theta}$	SD of θ_t	21.180 (3.422)	(14.473, 27.888)
$\gamma_{\omega,\omega}$	SD of ω_t	0.205 (0.049)	(0.110, 0.300)
$\phi_{\alpha,\alpha}$	Autocorrelation of α_t	0.334 (0.122)	(0.094, 0.574)
$\phi_{\theta,\theta}$	Autocorrelation of θ_t	0.524 (0.110)	(0.310, 0.739)
σ_α	SD of errors of α_t	0.166 (0.023)	(0.120, 0.212)
σ_θ	SD of errors of θ_t	18.034 (2.808)	(12.531, 23.538)
σ_ω	SD of errors of ω_t	0.205 (0.049)	(0.110, 0.300)
$\rho_{\alpha,\theta}$	Correlation between the errors of α_t and θ_t	-0.592 (0.125)	(-0.837, -0.346)
$\rho_{\alpha,\omega}$	Correlation between the errors of α_t and ω_t	-0.357 (0.287)	(-0.920, 0.206)
$\rho_{\theta,\omega}$	Correlation between the errors of θ_t and ω_t	-0.307 (0.254)	(-0.806, 0.191)
r_w^m	SD of mother effect	0.041 (0.013)	(0.015, 0.066)

FIGURE 2 Position of optimal laying date over study period from 1955 to 2015. The estimated movement of optimal laying date from the selected model is shown with solid blue line, along with its 95% confidence interval (dashed blue lines). The black dots are the observed within-year mean laying dates



model reported a significant and larger estimate of $\phi_{\theta,\theta}$ with narrower confidence interval (0.310, 0.739). With the result from the simulation study, even this larger estimate may be potentially underestimated. The estimates of the standard deviations of errors of VAR(1) α_t , θ_t and ω_t (σ_α , σ_θ , σ_ω) are also listed in Table 3, along with estimates of correlations of the correlated noises. Even though our result indicates that the VAR(1) stochastic processes α_t , θ_t and ω_t are correlated through errors w_t not through transition matrix Φ , we are conservative in interpreting the estimates of $\rho_{\alpha,\alpha}$ and $\rho_{\theta,\theta}$ since the candidate model 13 with only $\rho_{\alpha,\theta}$ reported almost the same AIC value as our best model. The estimate of standard deviation of mother effect is 0.041, implying that the mean of X_i from broods produced by the same mother are weakly correlated to each other.

The estimated optimum phenotype is shown in Figure 2 with a solid blue line. It fluctuates over the study period with an obvious downward trend. The mean within-year laying dates (denoted with black dots) also show a downward trend but the advance is not as strong as the optimum, resulting in increasing mistiming between the optimal laying date and the mean within-year laying date, which is in line with the finding from previous study of the same population (e.g. Chevin et al., 2015; Reed et al., 2013; Visser & Both, 2005; Visser et al., 1998). Since the reproductive fitness of the great tits depends strongly on the mismatch with food phenology, mistiming in our case therefore equals mismatch, even though the food resource phenology is not considered in our study (see Visser and Gienapp (2019) for the difference between mistiming and mismatch). One explanation for the mismatch is that females might be unwilling to breed at the optimal date in terms of the offspring fitness because of higher energetic cost of producing and incubating earlier in harsh environment where it is cold and food is scarce, mismatching by a few days might therefore be optimal for the sake of parental fitness (Te Marvelde, Webber, Meijer & Visser, 2011). Beside this optimal mismatch hypothesis, another leading explanation (the cues hypothesis) is that the cues used for timing laying are no longer accurately predicting the phenology of the food peak (see Visser et al., 2012 for more details on these two hypotheses).

3.4 | Model evaluation

The performance of our selected model is evaluated by visualizing the observed and predicted number of fledglings for each year. Each panel in Figure 3 shows the observed (dots in the panels) and predicted number of fledglings (dark solid line) against laying date for

a specified year (from 1955 to 2015). Our analysed data includes three brood types represented by three colours in the plots. The red, green and blue dots correspond to first, replacement and second broods respectively. The solid grey curve represents nonparametric loess regression through the points with the dashed grey lines being associated 95% confidence band. With our selected model,

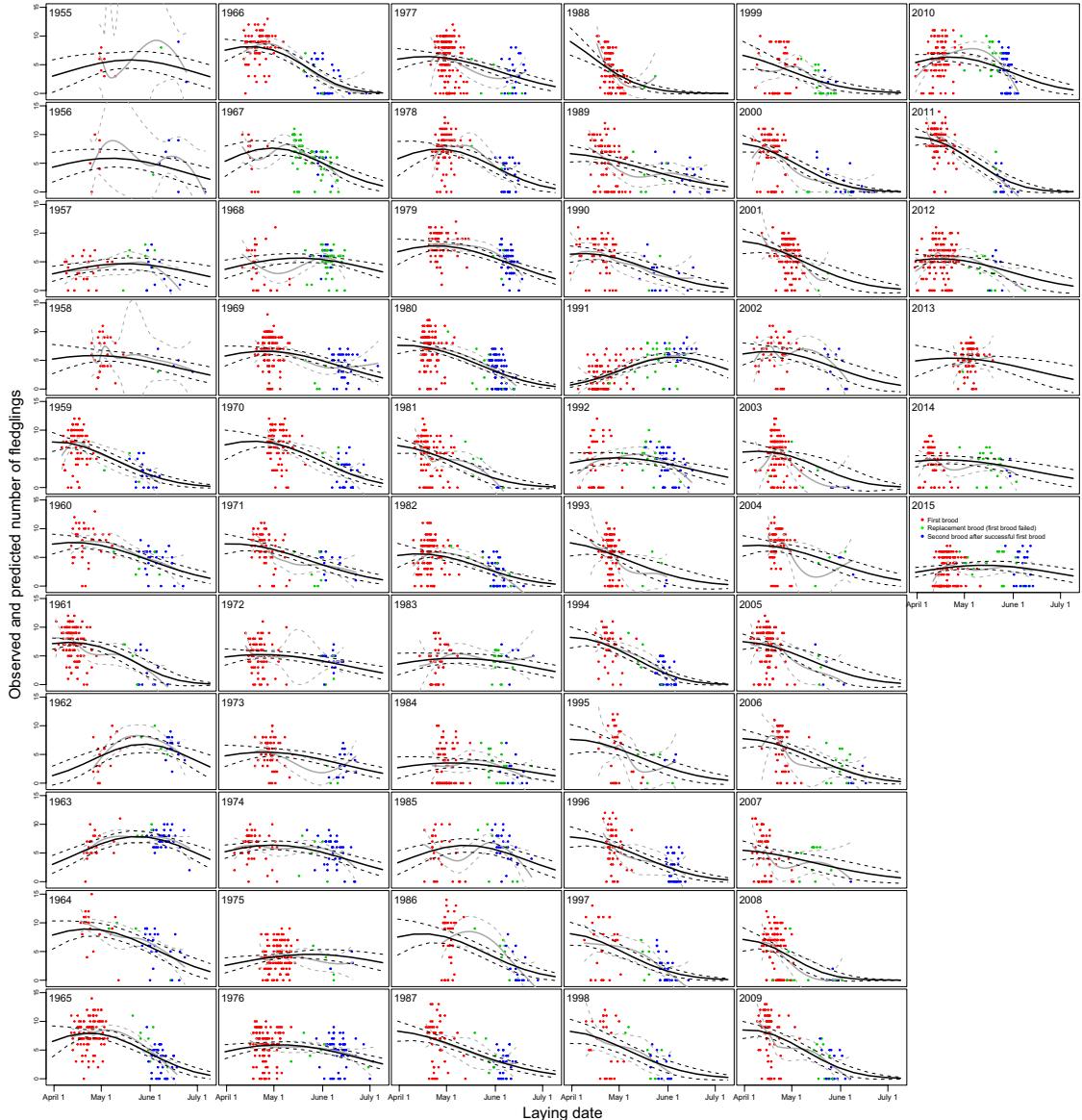


FIGURE 3 Observed and predicted number of fledglings ($E(Y_i|p_i, w_i)$) against the laying date for each year. The blue, red and green dots represent the observed number of fledglings from first, replacement and second broods respectively. The grey curve is loess regression (with default degree of smoothing = 0.75) through the scatter points with 95% confidence band (dashed grey lines). The black line indicates the number of fledglings predicted with our selected model conditional on zero mother random effects with dashed black lines representing its 95% confidence band. The 95% confidence band was calculated by multiplying the standard errors reported with TMB by the 2.5 and 97.5th percentiles of the normal distribution

the dark curve shows the predicted number of fledglings at laying dates conditional on zero mother effect over the whole breeding season, with the dashed dark lines representing associated 95% confidence band again. The figure indicates a good fit of our selected model to the data as we can see that the dark lines lie within the loess confidence bands for all the years. For most of the years, the prediction of number of fledglings peaked at early breeding season when the first broods were laid except year 1991, when a late frost hit the population and the plot validates our expectation of a very late optimum estimate.

4 | CONCLUSION AND POSSIBLE EXTENSIONS

Thanks to the new techniques such as TMB for fast likelihood computation for non-Gaussian and nonlinear models, the use of state-space models for analysing ecological systems is increasing (for example Cadigan, 2015; Albertsen, Nielsen & Thygesen, 2016; Auger-Méthé et al., 2017). The conditional independence structure in state-space models yield a sparse precision matrix for the joint distribution of the data and the random effects (Kristensen et al., 2015) and TMB takes maximal advantage of this sparseness structure (through automatic sparsity detection) in its computation of the Laplace approximation. Therefore, using state-spaces models coupled with TMB makes estimating a large number of parameters and random effects which is usually the case in modelling complicated biological processes or ecological systems, possible. Compared with the models and approaches adopted by previous studies on fluctuating selection, our method based on SSM, GLMM and TMB has several advantages. First, state-space models allow us to explore two correlated fitness components simultaneously, instead of measuring different fitness components independently. Second, due to the flexibility of SSMs, parameters can be estimated efficiently with little computational effort. Third, the formulation of our theoretical models turns out to be more realistic to account for directional selection and non-Gaussian fitness residual, as GLMMs. Our results from the great tit case study partly agree with the findings from previous studies on the same population, and due to the VAR(1) formulation for the fitness parameters we could gain more in terms of the underlying patterns of the fluctuating selection. For the researchers who are interested in applying our method to their data either for modelling fluctuating selection or general ecological systems with VAR(1) stochastic processes, it is worth to mention that TMB has no built-in probability function for modelling VAR(1). Our study can serve as a template for this as well as for conducting model selection with TMB.

In our study, we treated fluctuations in properties of the Gaussian fitness function as a vector autoregressive process. In principle, our approach can also accommodate other autocovariance structures, such as vector autoregressive moving-average (ARMA) models (see Wei, 2006 for the definition). Besides, in our statistical model, the random mother effect $\epsilon_{j(i)}$ is assumed to be same for the two episodes

but vary in magnitude, which implies that a mother that is likely to have complete brood failure will be more likely to have a low number of fledglings (with correlation 1). To relax this assumption, the mother effects can be treated differently for the two episodes and assumed to be multivariate Gaussian distributed $(\epsilon_{p,j(i)}, \epsilon_{w,j(i)})^T \sim N(\mathbf{0}, \Sigma)$ with Σ being a covariance matrix. Furthermore, the number of fledglings is chosen as the selection component so that the estimates could be compared to those from Chevin et al. (2015), which claims that using the number of recruits may cause more uncertainty in estimates of parameters due to the much larger coefficient of variation in the number of recruits. However, in some previous studies (e.g. Reed et al., 2013), the fecundity component of fitness is measured as the number of recruits surviving to the next breeding season instead of the number of individuals surviving to become fledglings. As claimed by Naef-Daenzer and Grüebler (2016), using the number of fledglings as a proxy for fitness may be misleading in inference of evolutionary significance since reproductive success can be completely altered by many causal factors driving the adaptations which operate during the post-fledgling period, and thereby change the juveniles' fate from fledgling to independence. Therefore, it would be interesting to expand our model to incorporate both pre- and post-fledgling period, such as chicks' survival and recruitment probability, as well as a female's survival, into a comprehensive life-history framework for the lifetime selection exploration.

Our study demonstrates a technique of estimating fluctuating selection in cases where ecological covariates are not available. To understand whether observed shift in selection are biologically meaningful, however, it is important to elucidate the ecological drivers of fluctuations in selection. Empirical investigations of the causal mechanisms driving such selection dynamism are needed before the development of novel analytical and statistical techniques. In our great tit case, for example, the peak movement might be affected by the height of the caterpillar peak, the mean breeding timing relative to the caterpillar peak and the breeding density. The width of the fitness function is likely being affected by the height and probably the width of the caterpillar peak (Visser et al., 2006). The location of optimum might be influenced by environmental variables (e.g. Chevin et al., 2015; Gienapp et al., 2013). However, the biotic interactions coupled with other abiotic factors playing a direct or indirect role in the selective process could likely make analysis much more complicated. Other extensions include analysis of correlational selection on multivariate traits and estimating the temporal-spatial variation and correlation in fluctuating selection.

ACKNOWLEDGEMENTS

This work was supported by the Research Council of Norway through its Centres of Excellence funding scheme (project number 223257 to CBD).

AUTHORS CONTRIBUTIONS

M.E.V. provided the data; J.T. conceived the idea and initiated the statistical model; Y.C. analysed the data and conducted the analyses;

Y.C. wrote the initial draft with input from J.T.; all authors contributed to revisions on later manuscript versions and gave final approval for publication.

DATA AVAILABILITY STATEMENT

All the necessary data and source code to carry out the analyses in this study are available in the Dryad Digital Repository <https://doi.org/10.5061/dryad.q4q8r89> (Cao, Visser & Tufto, 2019).

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd international symposium on information theory*, Akadémiai Kiadó, Budapest, 1973 (pp. 267–281).
- Albertsen, C. M., Nielsen, A., & Thygesen, U. H. (2016). Choosing the observational likelihood in state-space stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(5), 779–789.
- Auger-Méthé, M., Albertsen, C. M., Jonsen, I. D., Derocher, A. E., Lidgard, D. C., Studholme, K. R., ... Flemming, J. M. (2017). Spatiotemporal modelling of marine movement data using template model builder (TMB). *Marine Ecology Progress Series*, 565, 237–249.
- Barton, N. H., & Keightley, P. D. (2002). Multifactorial genetics: understanding quantitative genetic variation. *Nature Reviews Genetics*, 3(1), 11.
- Bell, G. (2010). Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537), 87–97.
- Bull, J. (1987). Evolution of phenotypic variance. *Evolution*, 41(2), 303–315.
- Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Berlin: Springer Science & Business Media.
- Cadigan, N. G. (2015). A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(2), 296–308.
- Calsbeek, B. (2011). Exploring variation in fitness surfaces over time or space. *Evolution*, 66(4), 1126–1137.
- Cao, Y. H., Visser, M. E., & Tufto, J. (2019). Data from: A time series model for estimating temporal variation in phenotypic selection on laying dates in a Dutch great tit population. *Dryad Digital Repository*, <https://doi.org/10.5061/dryad.q4q8r89>
- Chevin, L.-M., Visser, M. E., & Tufto, J. (2015). Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. *Evolution*, 69(9), 2319–2332.
- Cohen, D. (1966). Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, 12(1), 119–129.
- Futuyma, D. J. (2006). *Evolutionary biology* (3rd ed.). New York: W. H. Freeman; Basingstoke: Palgrave [distributor].
- Gamelon, M., Tufto, J., Nilsson, A. L. K., Jerstad, K., Røstad, O. W., Stenseth, N. C., ... Sæther, B. E. (2018). Environmental drivers of varying selective optima in a small passerine: a multivariate, multi-periodic approach. *Evolution*, 72(11), 2325–2342.
- Gienapp, P., Lof, M., Reed, T. E., McNamara, J., Verhulst, S., & Visser, M. E. (2013). Predicting demographically sustainable rates of adaptation: can great tit breeding time keep pace with climate change? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368 (1610).
- Gienapp, P., Reed, T. E., Visser, M. E. (2014). Why climate change will invariably alter selection pressures on phenology. *Proceedings of the Royal Society B: Biological Sciences* 281(1793), 20141611.
- Hedrick, P. W. (2006). Genetic polymorphism in heterogeneous environments: the age of genomics. *Annual Review of Ecology, Evolution, and Systematics*, 37(1), 67–93.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2015). TMB: automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660*.
- Lande, R., & Shannon, S. (1996). The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution*, 50(1), 434–437.
- Maziarz, M., Wesolowski, T., Hebda, G., Cholewa, M., & Broughton, R. K. (2016). Breeding success of the Great Tit *Parus major* in relation to attributes of natural nest cavities in a primeval forest. *Journal of Ornithology*, 157(1), 343–354.
- Morrissey, M. B., & Hadfield, J. D. (2012). Directional selection in temporally replicated studies is remarkably consistent. *Evolution*, 66(2), 435–442.
- Naef-Daenzer, B., & Grüebler, M. U. (2016). Post-fledging survival of altricial birds: ecological determinants and adaptation. *Journal of Field Ornithology*, 87(3), 227–250.
- Perrins, C. M. (1965). Population fluctuations and clutch-size in the Great Tit, *Parus major* L. *The Journal of Animal Ecology*, 601–647.
- Perrins, C. M. (1979). *British tits* (Vol. 62). New York, NY: HarperCollins.
- Post, E., & Forchhammer, M. C. (2007). Climate change reduces reproductive success of an arctic herbivore through trophic mismatch. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1501), 2367–2373.
- Reed, T. E., Jenouvrier, S., & Visser, M. E. (2013). Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology*, 82(1), 131–144.
- Revell, L. J. (2007). The G matrix under fluctuating correlational mutation and selection. *Evolution: International Journal of Organic Evolution*, 61(8), 1857–1872.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal Of The Royal Statistical Society Series B: (Statistical Methodology)*, 71(2), 319–392.
- Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity. *Annual Review of Ecology and Systematics*, 24(1), 35–68.
- Sieminski, A. M., DiBattista, J. D., & Carlson, S. M. (2009). It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecology Letters*, 12(11), 1261–1276.
- Svardal, H., Rueffler, C., & Hermisson, J. (2011). Comparing environmental and genetic variance as adaptive response to fluctuating selection. *Evolution*, 65(9), 2492–2513.
- Te Marvelde, L., Webber, S. L., Meijer, H. A. J., & Visser, M. E. (2011). Mismatched reproduction is energetically costly for chick feeding female great tits. *Functional Ecology*, 25(6), 1302–1308.
- Townsend, A. K., Sillett, T. S., Lany, N. K., Kaiser, S. A., Rodenhouse, N. L., Webster, M. S., Holmes, R. T. (2013). Warm springs, early lay dates, and double brooding in a North American migratory songbird, the black-throated blue warbler. *PLoS ONE*, 8(4), 59467.
- Tufto, J. (2015). Genetic evolution, plasticity, and bet-hedging as adaptive responses to temporally autocorrelated fluctuating selection: a quantitative genetic model. *Evolution*, 69(8), 2034–2049.
- Van Tienderen, P. H., & Koelewijn, H. P. (1994). Selection on reaction norms, genetic correlations and constraints. *Genetics Research*, 64 (2), 115–125.
- Visser, M. E., & Both, C. (2005). Shifts in phenology due to global climate change: the need for a yardstick. *Proceedings of the Royal Society B: Biological Sciences*, 272(1581), 2561–2569.
- Visser, M. E., & Gienapp, P. (2019). Evolutionary and demographic consequences of phenological mismatches. *Nature Ecology and Evolution*, 3, 879–885.
- Visser, M. E., Gienapp, P., Husby, A., Morrissey, M., de la Hera, I., Pulido, F., Both, C. (2015). Effects of spring temperatures on the strength of

- selection on timing of reproduction in a long-distance migratory bird. *PLoS Biology*, 13 (4), 1–17.
- Visser, M. E., Holleman, L. J., & Gienapp, P. (2006). Shifts in caterpillar biomass phenology due to climate change and its impact on the breeding biology of an insectivorous bird. *Oecologia*, 147 (1), 164–172.
- Visser, M. E., Marrelde, L. te, & Lof, M. E. (2012). Adaptive phenological mismatches of birds and their food in a warming world. *Journal of Ornithology*, 153, 75–84.
- Visser, M. E., Noordwijk, A.V., Tinbergen, J. M., Lessells, C. M. (1998). Warmer springs lead to mistimes reproduction in Great Tits (*Parus major*). *Proceedings of the Royal Society B*, 265(1408), 1867–1870.
- Wei, W. (2006). Time series analysis: Univariate and multivariate methods. (2 ed.). Boston, MA: Pearson Addison Wesley.
- Zhang, X. S., & Hill, W. G. (2005). Evolution of the environmental component of the phenotypic variance: stabilizing selection in changing environments and the cost of homogeneity. *Evolution*, 59(6), 1237–1244.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Cao Y, Visser ME, Tufto J. A time series model for estimating temporal variation in phenotypic selection on laying dates in a Dutch great tit population. *Methods Ecol Evol*. 2019;10:1401–1411. <https://doi.org/10.1111/2041-210X.13249>

Supporting Information (SI) for

A time series model for estimating temporal variation in phenotypic selection on laying dates in a Dutch great tit population

About TMB

Template Model Builder (TMB; Kristensen, Nielsen, Berg, Skaug, & Bell, 2015)) is an R package for fitting statistical latent variable models. It is functionally similar to ADMB (Fournier et al., 2012). The joint likelihood for the data and the random effects are defined by the user as a C++ template function. Then the package evaluates and maximizes the Laplace approximation of the marginal likelihood where the random effects are automatically integrated out. This approximation is achieved by using reverse-mode automatic differentiation (up to order three) of the joint likelihood. The combination of reverse-mode automatic differentiation and the Laplace approximation for high-dimension integrals allows for the efficient fitting of complex (nonlinear, non-Gaussian, and hierarchical) models with large multivariate data sets to perform parameter estimation (Fournier et al., 2012). TMB takes maximal advantage of sparseness structure (Kristensen et al., 2015) and the first derivatives of the Laplace approximation obtained with automatic differentiation of the negative log-likelihood can be used by other approaches such as hybrid MCMC.

More details on study population

The great tit is 18–20g small passerine bird species widespread throughout European woodlands and gardens. As a cavity nester, it readily accepts nest-boxes for breeding, which allows monitoring of the whole population if a surplus of nest-boxes is provided (Harvey, Greenwood, & Perrins, 1979). The study area consists of mixed pine-deciduous woodland on poor sandy soils. From 1955 to 2015, more nest boxes than required were placed in the study area at approximately constant availability. On average the ratio of nest boxes to breeding females was around 3 : 1 in a typical year. During the breeding season from April to June/July, nest boxes were visited once per week. At each visit, the number of eggs or nestlings was counted and nestlings were given metal leg rings on day 7 and the parents caught on the nest using a spring trap. For some years clutch or brood size manipulation experiments were carried out, which possibly affected fledgling production or recruitment probability, therefore, manipulated broods were excluded from our analysis. We also deleted 35 third clutch observations for simplifying the comparison between the different brood types. We also deleted the records with uncertainty of the brood type, and clutch size being smaller than number of fledglings. Unknown females were not included in our analyses, as their mother effects as random effects in the model could not be determined. Eventually, 5892 out of 6353 records were kept for our analysis.

Simulation study

A simulation study was carried out to test the power of our method in identifying the best model. We simulated laying dates z which stabilizing selection acts on with a mixture of normal distribution $0.7N(23, 7.5) + 0.3N(62, 10.5)$, which is close to the reality of the Dutch great tit population. We considered 50 years and for each year the sample size was drawn from a Poisson distribution with a mean of $n = 100$ individuals. To simplify the simulation study, we considered stabilizing selection via the expected number of fledglings (episode W) while the zero-inflated probability (episode P) was kept as a fixed parameter (set to 0.12). In terms of the parameters in equation (3) and (4) in the main text, vector $(\bar{\alpha}, \bar{\theta}, \bar{\omega})$ was set to $(2, 18, \log(45))$ and the vector of standard deviation of the random effects $(\sigma_\alpha, \sigma_\theta, \sigma_\omega)$ was set to $(0.2, 18, 0.2)$. For brevity the variance-covariance matrix Σ was set diagonal and only $\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$ in the transition matrix Φ were considered as non-zero. They were set to be equal ($\phi_{\alpha,\alpha} = \phi_{\theta,\theta}$) and took values from $(0, 0.1, 0.25, 0.5, 0.75, 0.9)$.

The standard deviation of random mother effects was set to 0.05 and added only to episode W .

Since there is a potentially long list of candidate models, we did not fit all the possible models and instead considered, in addition to the true model, five models that can help us to test if our model selection procedure has the power to: identify the zero-inflation probability as a parameter of a selection episode; distinguish models with and without fluctuation in ω_t ; identify the auto-correlation parameters in Φ ; distinguish models with and without off-diagonal parameters in Φ ; identify random mother effects in the ‘true’ model. Specifically, based on the true model we fitted (i) a model with zero-inflation probability regressed against laying dates with random intercepts but without random slopes; (ii) a model with fixed ω_t ; (iii) a model with all entries in Φ equal to zero; (iv) a model with 2×2 upper-left non-zero entries in Φ ; (v) a model without random mother effects. For each value of $\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$, we ran 100 simulations and for each simulation we compared the reported AIC between the true model and each of the alternative models respectively. The true model was selected against the alternative model only when the AIC of it is at least two points lower than that of the alternative model.

Fig. S1 shows the simulation result. The left plot shows the percentage of cases for which the true model was selected over each of the alternative models against the actual auto-correlation values used in the simulations. It is clear that when zero-inflation probability is only a fixed parameter in the model, the model selection procedure never wrongly favours the model with fluctuating zero-inflation probability. When ω_t was set to fixed in an alternative model, the true model with random ω_t was detected in approximately 80% of the simulations. If $\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$ were excluded from the true model, it then reduced to an alternative model with α_t , θ_t and ω_t following iid processes. The true model (including auto-correlation) is rarely selected as best over the alternative model when the auto-correlation value is as small as 0.1. However, when the

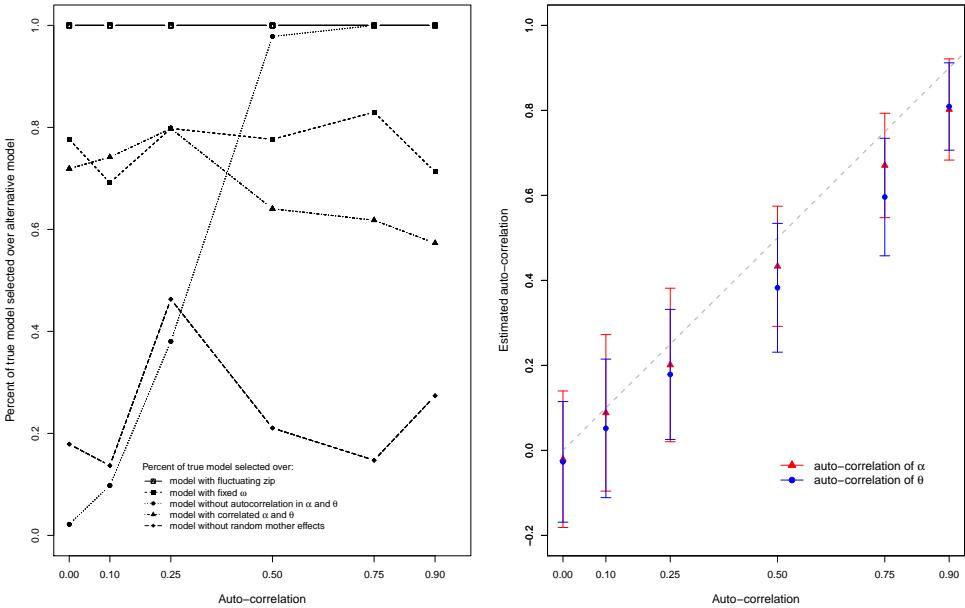


Figure S1: Left: test the power of our method. Each of the line in the plot shows the percentage of our true model was selected against an alternative model out of total 100 simulations. The x axis is the actual values of $\phi_{\alpha,\alpha} = \phi_{\theta,\theta}$ used in simulations. The five lines represent five alternative models that the true model were compared with and the true model was chosen by at least two points lower in AIC than the alternative model. The five alternative models are described in the legend. Right: estimated auto-correlation ($\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$) in all the simulations. The red triangles represent the mean of $\hat{\phi}_{\alpha,\alpha}$ under each setting of the auto-correlation, with error bars representing ± 1 standard deviation of the estimates over 100 simulations. The similar explanation applies to $\hat{\phi}_{\theta,\theta}$, which is shown with blue color. The dashed grey line plots the expected value if the MLEs are unbiased.

auto-correlation magnitude increased to 0.5, in over 90% of the simulations the AR(1) structure in α_t and θ_t can be detected. When cross-correlation between α_t and θ_t ($\phi_{\alpha,\theta}$, $\phi_{\theta,\alpha}$) was added to the true model, only in around 60% of the simulations was the true model selected against the more complex alternative model. It is even more challenging when the random mother effects were excluded from the true model, that in only around 20% of the simulations that the true model with random mother effects was chosen.

The right plot in Fig. S1 shows the estimated auto-correlation against the true auto-correlation used in the simulations. Red and blue color corresponds to $\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$ respectively. The triangles and round dots show the mean estimate of $\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$ respectively with error bars representing one standard deviation of all the estimates over 100 simulations. The dashed grey line represents the expected value if the MLEs are unbiased and it goes across all the error bars of $\hat{\phi}_{\alpha,\alpha}$. Estimating $\phi_{\theta,\theta}$ accurately turns more difficult than that for $\phi_{\alpha,\alpha}$ indicated by the larger deviation from the unbiased MLEs.

To sum up, our model fitting and model selection procedure has promising power to capture the basic structure (fixed zero-inflation probability, fluctuated ω_t and auto-correlation in α_t and θ_t) of the true model. However, it also shows that the cross-correlation between the fitness parameters might be overestimated and therefore wrongly included in the selected model. In addition, cautions should be made when excluding random mother effects from the model especially when the models with and without random mother effects report similar AIC values since the random mother effects might be too small to be detected with AIC. At last, both our simulation study and the one in Chevin, Visser, and Tufte (2015) shows that the auto-correlations in the fitness parameters are potentially underestimated.

Supplementary model selection procedure

In the main text we have shown the model selection procedure only for the best model and the models around it. The updating procedure from a null model to the best model is supplemented in Table S1. Model 1 is consistent with stabilizing selection via episode W and episode P and the fitness function parameters remain unchanged across years but vary across episode W and P . Based on the estimates of the parameters in model 1, we changed stabilizing selection into directional selection for episode W (model 2, 2a, 2b, 2c) or for episode P (model 3, 3a, 3b, 3c). The models with correlated random intercepts and slopes (model 2c and 3c) perform best in each situation. Next, from model 4 to model 8b, we updated each model (model 1, 2c, 3c) such that α_t , θ_t and ω_t are either white noise (model 4 to model 4e, note that model 4a, 4c, 4e are hard to get converged, thus they were updated through model 4, 4b and 4d respectively), or AR(1) (model 5 to model 5b), or VAR(1) (model 6 to model 6b) processes. Model 7, 7a, 7b, 7c were updated with only significant entries of Φ kept. Auto-correlations between errors of α_t , θ_t and ω_t were introduced into model

8, 8a and 8b. So far, model 8 reports the smallest AIC, therefore, mother effect was added to it (model 9), and model 9 was eventually confirmed to be the best model, as have shown in the main text.

It is worth to mention that our candidate models were generally updated from the simple ones to the complicated ones and therefore the subsequent models are subject to the choice of the initial models. The choice should be made carefully especially when the initial models report similar AIC values. In this case, one suggestion is that the subsequent models can be updated simultaneously based on the competitive initial models and another suggestion is the neighbor models of the selected model should be carefully tested to ensure it is indeed the best one.

Table S1: Supplementary Model selection procedure of phenotypic selection on breeding time of great tits. The order of models listed below is accordance with the order of models fitting, from model 1 to 9. The following model selection procedure can be found in the main text. ΔAIC and Δp is the difference in AIC and number of parameters p between each model and the best model (model 9). The column of description gives the details of updating model based on the previous ones. For simplification, the probability of successful-brooding component is denoted as episode P and the mean number of fledglings as episode W .

Model	ΔAIC	Δp	Description
1	1246.99	-12	$\eta_{s,t}^{(\alpha)}, \eta_{s,t}^{(\theta)}, \eta_{s,t}^{(\omega)}$ fixed across t , vary across s
based on model 1, change stabilizing selection via episode P into directional selection			
2	1246.01	-13	only with fixed intercept and slope
2a	1011.41	-12	add random intercepts on model 2
2b	919.21	-11	add random slopes on model 2a
2c	877.37	-10	add covariance to random intercepts and slopes on model 2b
based on model 1, change stabilizing selection via episode W into directional selection			
3	1247.54	-13	only with fixed intercept and slope
3a	708.48	-12	add random intercepts on model 3
3b	509.98	-11	add random slopes on model 3a
3c	437.33	-10	add covariance to random intercepts and slopes on model 3b
change fixed α_t, θ_t and ω_t into white noise			
4	241.24	-10	based on model 1, white noise α_t and ω_t , fixed θ_t
4a	114.45	-9	based on model 4, white noise α_t, θ_t and ω_t
4b	114.08	-8	based on model 2c, white noise α_t and ω_t , fixed θ_t
4c	110.11	-7	based on model 4b, white noise α_t, θ_t and ω_t
4d	129.33	-8	based on model 3c, white noise α_t and ω_t , fixed θ_t

Continued on next page

Table S1 – continued from previous page

Model	ΔAIC	Δp	Description
4e	96.48	-7	based on model 4d, white noise α_t , θ_t and ω_t
change random α_t , θ_t and ω_t into AR(1)			
5	81.74	-6	based on model 4a, AR(1) α_t , θ_t and ω_t
5a	40.84	-4	based on model 4c, AR(1) α_t , θ_t and ω_t
5b	74.57	-4	based on model 4e, AR(1) α_t , θ_t and ω_t
change random α_t , θ_t and ω_t into VAR(1)			
6	89.02	0	based on model 4a, VAR(1) α_t , θ_t and ω_t
6a	88.83	2	based on model 4c, VAR(1) α_t , θ_t and ω_t
6b	80.59	1	based on model 4e, VAR(1) α_t , θ_t and ω_t
keep only significant (at significance statistics 0.05) entries in Φ			
7	82.17	-7	update based on model 6, AR(1) α_t and θ_t
7a	39.13	-5	no significant entries in Φ in model 6a, so update based on model 5a, AR(1) α_t and θ_t
7b	76.48	-5	update based on model 6b, AR(1) α_t and θ_t
7c	76.1	-6	update based on model 6b, AR(1) α_t
add correlations to the errors of α_t , θ_t and ω_t			
8	57.8	-3	update based on model 5, which is the best model so far for stabilizing selection for both episode P and W
8a	19.1	-2	update based on model 7a, which is the best model so far for directional selection via P
8b	74.82	-1	update based on model 5b, which is the best model so far for directional selection via episode W
add mother effect			
9	0	0	update based on model 8a, which is the best model so far (directional selection via episode P and stabilizing selection via episode W)

Supplementary model evaluation

In addition to the model evaluation in the main text, we here further illustrate the performance of our selected model in predicting successful-brooding indices and non-zero number of fledglings. Each panel in Fig. S2 shows the observed indices and the predicted probability of successful-brooding against laying date for a specified year (from 1955 to 2015). Our analyzed data includes three brood types. The red, green and blue dots represent the observed indices of successful-brooding for first broods, replacements broods (first broods failed) and second broods (first broods succeeded) respectively. The solid grey curve represents nonparametric loess regression through the points with the dashed grey lines being associated 95% confidence band. The dark curve shows the predicted probability of successful-brooding at laying dates along the whole breeding season and conditional on zero random mother effects. It can be shown that it is a function of both the zero inflation probability and the mean number of fledglings:

$$\begin{aligned} P(Y_i > 0 | p_i, w_i) &= 1 - P(Y_i = 0 | p_i, w_i) \\ &= 1 - P(I_i = 0 | p_i) - P(X_i = 0 | I_i = 1, w_i)P(I_i = 1 | p_i) \\ &= (1 - p_i)(1 - e^{-w_i}), \end{aligned} \quad (\text{S1})$$

where p_i and w_i are zero-inflation probability and mean number of fledglings for brood i and estimated with our selected model. Similarly, each panel in Fig. S3 shows the observed number of fledglings (only nonzero observations plotted) and the expected number of fledglings predicted with our selected model for each year. The dots with different colors illustrate the observed number of fledglings from three brood types and the grey line again indicates the nonparametric loess regression with its 95% confidence band (dashed grey lines). The dark curve is the conditional expectation of number of fledglings ($E(Y_i | Y_i > 0)$) as a function of w_i estimated with our selected model with associated 95% confidence band (dashed black lines). Specifically, using the law of total expectation, we know that

$$E(Y_i | p_i, w_i) = E(Y_i | Y_i > 0, w_i)P(Y_i > 0 | p_i, w_i) + E(Y_i | Y_i = 0, p_i, w_i)P(Y_i = 0 | p_i, w_i).$$

Hence

$$\begin{aligned} E(Y_i | Y_i > 0, w_i) &= \frac{E(Y_i | p_i, w_i)}{P(Y_i > 0 | p_i, w_i)} \\ &= \frac{(1 - p_i)w_i}{(1 - p_i)(1 - e^{-w_i})} \\ &= \frac{w_i}{1 - e^{-w_i}}. \end{aligned} \quad (\text{S2})$$

It is worth noting that p_i and w_i in equation (S1) and equation (S2) were calculated conditional on zero

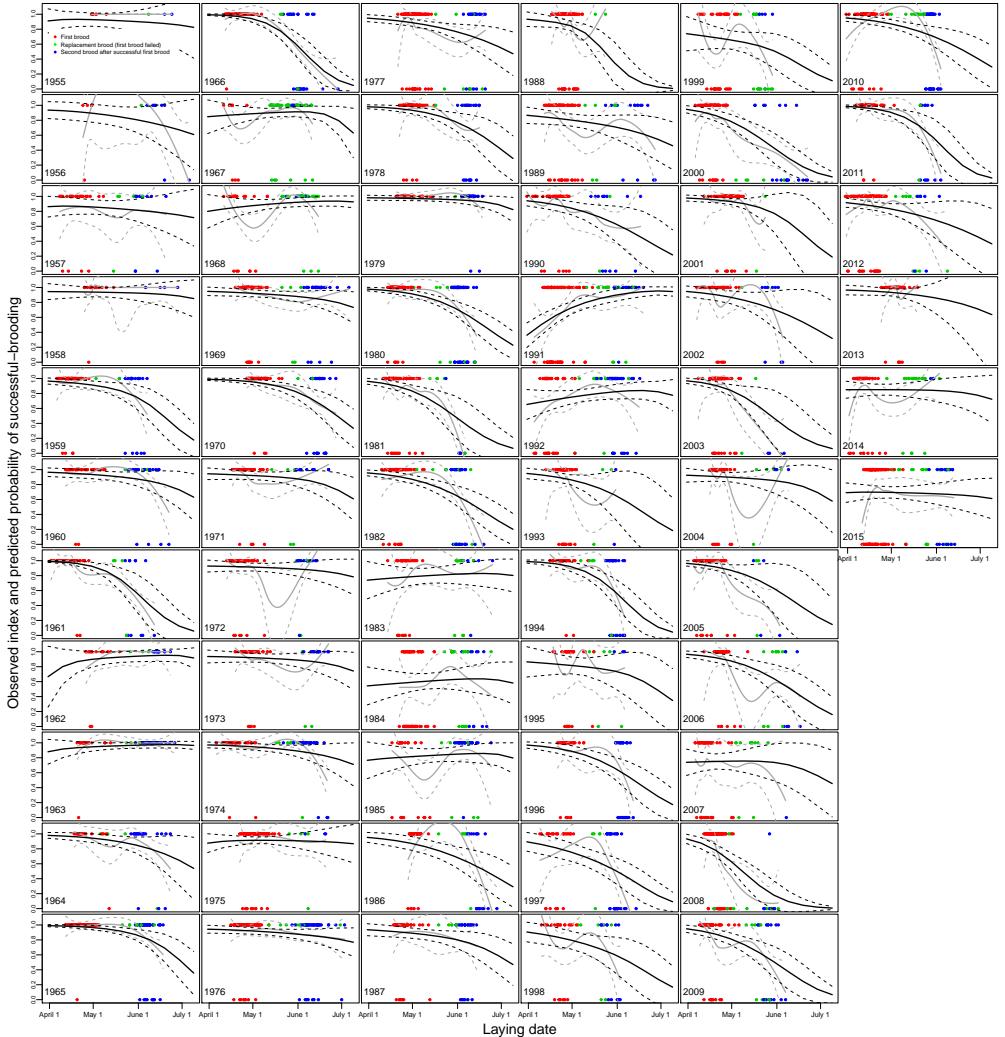


Figure S2: Observed indices and predicted probability of successful-brooding ($P(Y_i > 0 | p_i, w_i)$) against the laying date for each year. The blue, red and green dots represent the observed indices of successful-brooding for three different brood types (red dot represents first brood, green dot is replacement brood with first brood failed, and blue dot is second brood with successful first brood). The grey curve is loess regression (with default degree of smoothing = 0.75) through the scatter points with 95% confidence band (dashed grey lines). The black line indicates the probability of successful-brooding predicted with our selected model at laying dates along the whole breeding season and conditional on zero random mother effects and the dashed black lines represent associated 95% confidence band.

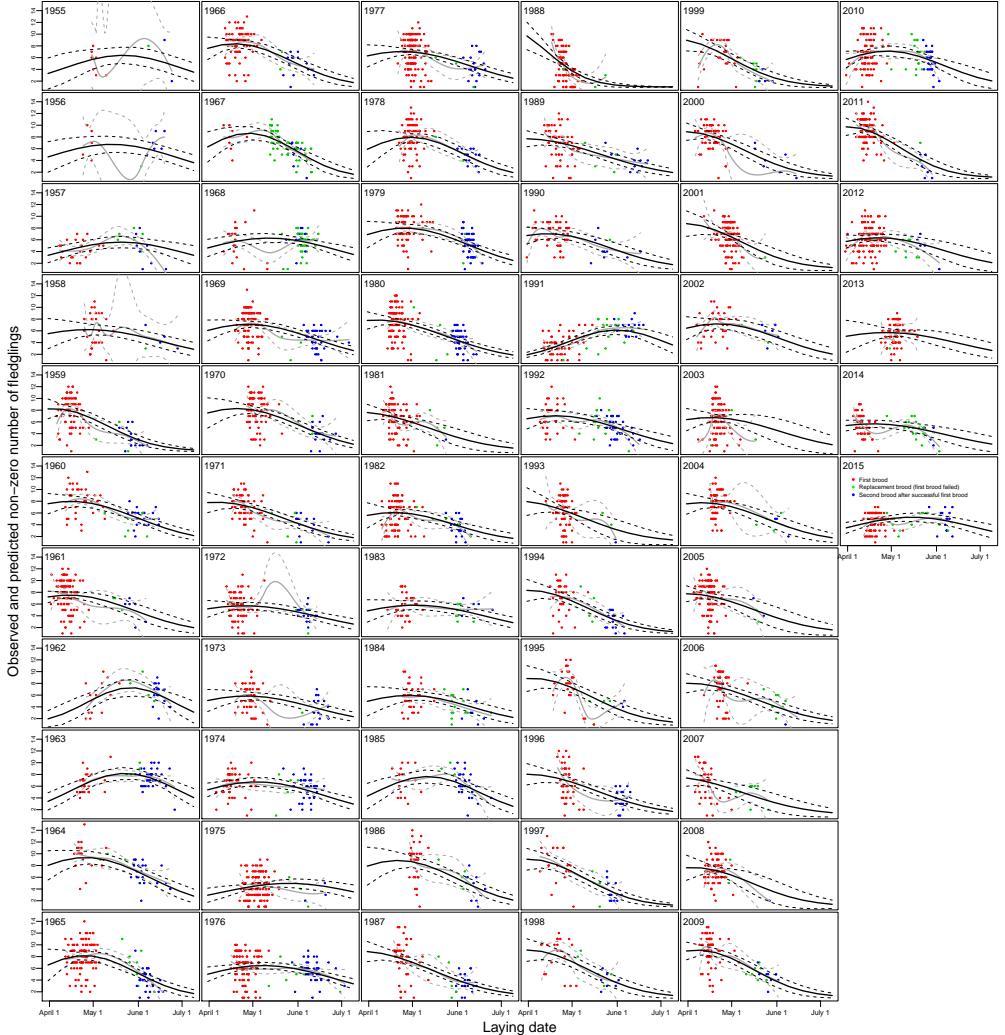


Figure S3: Observed and predicted nonzero number of fledglings ($E(Y_i | Y_i > 0, w_i)$) against the laying date for each year. Note that only nonzero number of fledglings are plotted. The blue, red and green dots represent the observed number of fledglings for three different brood types (red dot represents first brood, green dot is replacement brood with first brood failed, and blue dot is second brood with successful first brood). The grey curve is loess regression (with default degree of smoothing = 0.75) through the scatter points with 95% confidence band (dashed grey lines). The black line indicates the number of fledglings predicted with our selected model conditional on zero random mother effects with dashed black lines representing its 95% confidence band.

random mother effects for simplification. From both Fig. S2 and S3 we can see that for most of the years the dark line lies within the 95% confidence band of loess regression, indicating a good fit of our selected model. Moreover, we find from both figures that the replacement broods (first brood failed, green dots) were laid earlier than the second broods (first brood succeed, blue dots). Interestingly, it is hard to see any difference in the probability of successful-brooding between the replacement broods and second broods after successful first broods, but the mean number of fledglings for the second broods after successful first broods are strikingly smaller than that of the replacement broods. This might result from the increasing mistiming between the breeding time and optimal breeding time, and the fitness effects of being mismatched relative to the food peak are stronger at the individual level for the mean number of fledglings compared with the probability of successful-brooding. However, it is interesting to see that for most of the years the second broods suffer higher probability of complete loss than the first broods, as have been discussed in the main text.

Model fitting with partial data

In our study we used the great tit data of 1955-2015 (61 years) from the Hoge Veluwe. However, because a severe storm damaged the pine plantation in the winter of 1972-1973, some of the nest-boxes had to be replaced or relocated. Therefore, some of previous study on HV great tit population treated the HV1 (1955-1972) and HV2 (1973-2004) as two temporally separate populations (see Husby, Kruuk, & Visser, 2009). Other studies only focused on the HV great tit data after 1973 (for example Reed, Jenouvrier, & Visser, 2013; Gamelon et al., 2016). It is of our interest to fit the selected model with the data after 1973 and make a comparison between the estimates with this partial data set and full data set.

Table S2 shows the comparison between the estimates from our selected model with full data (1955-2015, 5892 records) and partial data (1973-2015, 4449 records), and the estimates from the model with fixed ω with partial data. We find from the second and third column that most of the estimates with the full data and with partial data are close to each other, but three differences are worth noting. First, the estimate of $\bar{\theta}$ is smaller (14.95 days) with partial data, which is reasonable and consistent with what can be seen from Fig. S4, where for recent years the estimated optimal laying dates are earlier compared with that in previous years. Second, the estimates of transition ($\phi_{\alpha,\alpha}$ and $\phi_{\theta,\theta}$) are slightly smaller with partial data. At last, the estimates with full data generally have less uncertainty (smaller estimate of standard error). Since the selected model from Chevin et al. (2015) assumed fixed ω_t across the study period from 1973 to 2015, it is interesting to get a flavor that how our result obtained from a candidate model with fixed ω_t and with data after 1973 differ from theirs. The last column in Table S2 therefore lists the estimates of

Table S2: Estimates (standard error) of model parameters from the selected model with full data and partial data, and the model with fixed ω_t and partial data.

Parameter	Estimate(S.E.)		
	Selected model		Model with fixed ω
	Full data	Partial data	Partial data
$\bar{\alpha}$	2.000(0.036)	1.996(0.041)	1.998(0.041)
$\bar{\theta}$	18.227(5.826)	14.950(5.753)	15.841(5.159)
$e^{\bar{\omega}}$ (days)	47.395(3.234)	45.985(3.835)	44.785(2.774)
$\gamma_{\alpha,\alpha}$	0.176(0.024)	0.181(0.031)	0.182(0.029)
$\gamma_{\theta,\theta}$	21.180 (3.422)	18.131(3.330)	19.423(2.838)
$\gamma_{\omega,\omega}$	0.205 (0.049)	0.191(0.056)	NA
$\phi_{\alpha,\alpha}$	0.334(0.122)	0.206(0.173)	0.251(0.161)
$\phi_{\theta,\theta}$	0.524(0.110)	0.386(0.157)	0.338(0.146)
σ_{α}	0.166(0.023)	0.177(0.032)	0.176(0.029)
σ_{θ}	18.034(2.808)	16.728(3.095)	18.278(2.694)
σ_{ω}	0.205(0.049)	0.191(0.056)	NA
$\beta_p^{(0)}$	2.946 (0.220)	2.742(0.233)	2.739(0.234)
$\beta_p^{(1)}$	-0.025(0.005)	-0.028(0.005)	-0.028(0.005)
σ_p^1	0.032(0.004)	0.028(0.004)	0.029(0.005)
ρ_p	-0.827(0.054)	-0.830(0.062)	-0.831(0.061)
β_v^m	0.701(0.092)	0.654(0.103)	0.653(0.103)
β_w^m	0.041(0.013)	0.041(0.016)	0.041(0.016)

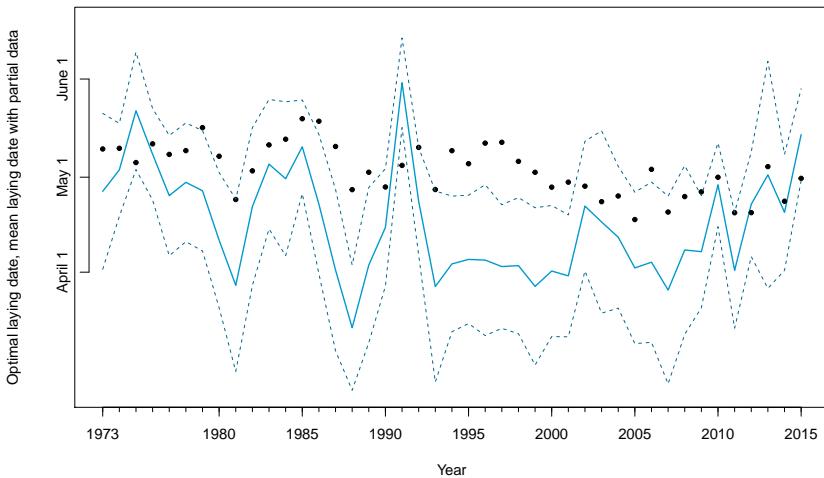


Figure S4: Position of optimal laying date estimated from our selected model with partial data from 1973 to 2015. The estimated movement of optimal laying date from the selected model is shown with solid blue line, along with its 95% confidence interval (dashed blue lines). The black dots are the observed within-year mean laying dates.

parameters in a model with the same formulation as our selected model except for that ω_t is assumed to be constant. We can see that the estimates in the last column are not far from that in the third column and the basic conclusion made from comparison with estimates from Chevin et al. (2015) remain the same, that our result reports larger width of fitness function, larger autocorrelation of the optimum laying dates and larger standard deviation of the fluctuated optimum laying dates.

Fig. S4 shows the movement of optimal laying date estimated from our selected model with partial data from 1973 to 2015. The estimated movement of optimal laying date from the selected model is shown with solid blue line, along with its 95% confidence interval (dashed blue lines). The black dots indicate the observed within-year mean laying dates. The pattern of the optimum movement in Fig. S4 is exactly identical to the movement of optimum in Fig. 2 in the main text from 1973 to 2015. This again implies that the full data set from 1955 to 2015 can be assumed to be generated from the same great tit population without invalidating the general results of our analysis.

Supplementary figures

The estimates of ω_t from our selected model range from 3.47 to 4.04 over the study period, and the corresponding natural exponent e^{ω_t} fluctuates from 32.15 to 56.65 days and the fluctuation can be seen clearly from the top-left plot of Fig. S5, even though the estimate of variance of ω_t is negligible and the candidate model with fixed ω_t does not perform much worse than our selected model. The movement of estimated within-year max fitness α_t (max mean number of fledglings), probability of successful-brooding and mean number of fledglings are shown in the top-right, bottom-left and bottom-right plot respectively, with the colorful lines representing non-parametric local regressions. 1988 is a standing-out year with a narrow width (32.49), large maximum number of fledglings (11.29) and early optimal laying date (14th March), which implies strong stabilizing selection via the mean number of fledglings (episode W) on laying dates. From the bottom plots the good years (1979, for example) with high mean probability of successful-brooding and mean number of fledglings can be differentiated from the bad years (1984, for example). The information obtained from the plots might provide insights for future researches which investigate the potentially abiotic variables driving the selection.

Although our approach offers an advance in the study of phenotypic selection, we believe we have not yet made best use of their full potential. Here, the estimates of random slopes produce order 1 autocorrelation in the annual directional selection as shown in Fig. S6. This implies that our model specification failed to capture the correlation structure of the fluctuated directional selection. While it does not bias the random slopes estimates, the standard deviation of the random slopes tends to be underestimated when the lag 1

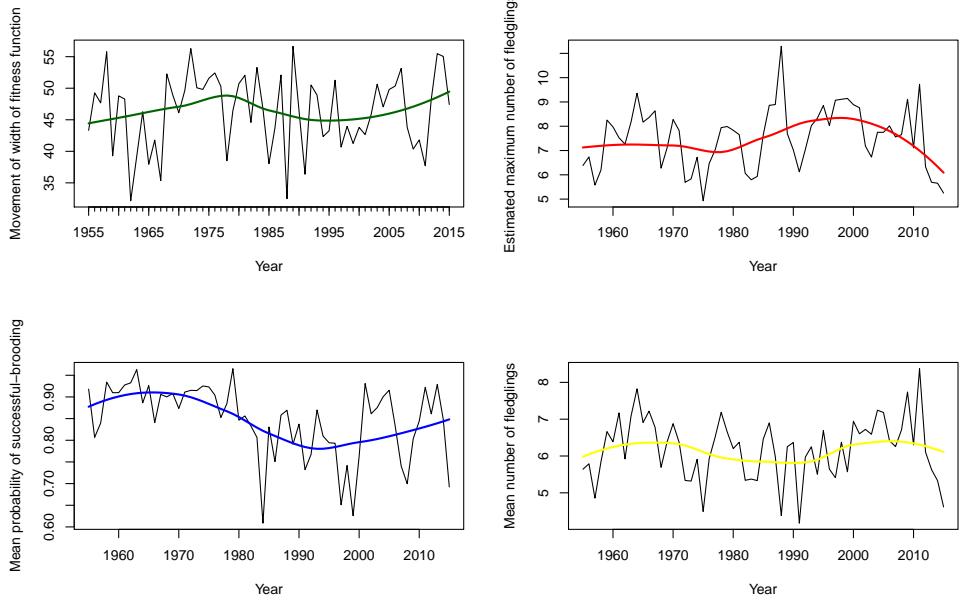


Figure S5: Annual movement of width of fitness function (top-left), maximum fitness (top-right), mean probability of successful-brooding (bottom-left) and mean number of fledglings (bottom-right). The black fluctuated lines are the corresponding estimates from our selected model (the discrete estimates are connected across years) and the colorful lines represent non-parametric local regressions.

autocorrelation of estimates is present.

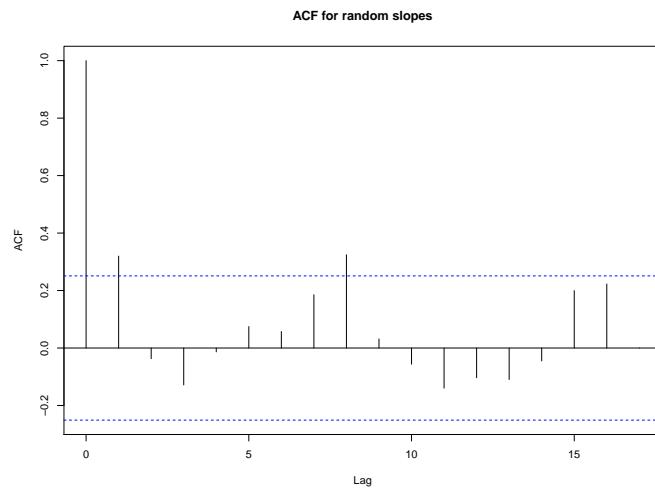


Figure S6: Estimated autocorrelation function of annual random slopes in the selected model for the probability of successful-brooding.

References

- Chevin, L.-M., Visser, M. E., & Tufto, J. (2015). Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. *Evolution*, 69(9), 2319–2332. doi: 10.1111/evo.12741
- Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., ... Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2), 233–249. doi: 10.1080/10556788.2011.597854
- Gamelon, M., Grøtan, V., Engen, S., Bjørkvoll, E., Visser, M. E., & Sæther, B. E. (2016). Density dependence in an age-structured population of great tits: Identifying the critical age classes. *Ecology*, 97(9), 2479–2490. doi: 10.1002/ecy.1442
- Harvey, P. H., Greenwood, P. J., & Perrins, C. M. (1979). Breeding Area Fidelity of Great Tits (*Parus major*). *The Journal of Animal Ecology*, 48(1), 305.
- Husby, A., Kruuk, L. E., & Visser, M. E. (2009). Decline in the frequency and benefits of multiple brooding in great tits as a consequence of a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, 276(1663), 1845–1854. doi: 10.1098/rspb.2008.1937
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2015). TMB: automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660*. doi: 10.18637/jss.v070.i05
- Reed, T. E., Jenouvrier, S., & Visser, M. E. (2013). Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology*, 82(1), 131–144. doi: 10.1111/j.1365-2656.2012.02020.x

Evolution of double brooding

Jarle Tufto^{*†}, Yihan Cao*, Marcel E. Visser[‡]

Thu Jan 30 11:24:02 2020

Abstract

In some populations of birds, females produce a second brood after raising a successful first brood. The proportion of females doing so varies strongly among study populations and years. To understand the adaptive significance of double brooding we consider double brooding jointly with the evolution of onset of breeding in a model with resources limited to a finite window in time. Double versus single brooding is modeled as a threshold character. Onset of breeding and the liability of double brooding follows a binormal phenotypic distribution. Depending on the cost of laying two broods versus one and the delay between the first and second brood relative to the width of the resource window and the phenotypic variance of onset of breeding, the adaptive topography may have single or multiple, purely single- or purely double-brooding adaptive peaks. Despite no frequency-dependence, an adaptive peak at an intermediate frequency of double brooding can exist if double brooding has a sufficiently negative phenotypic correlation with onset of breeding. If the location of the resource windows in time fluctuates between years, double-brooding has an additional adaptive value as a conservative bet-hedging strategy. Climate change, producing a linear trend in the location of the resource window towards earlier dates, may select for a reduced frequency of double brooding. An opposite effect is also possible if the additive genetic covariance between the liability and onset of breeding is negative. Finally, the model is discussed in terms of an empirical example.

1 Introduction

In many bird species, females produce a second brood within the same season after raising a successful first brood. Understanding the mechanism of multi-brooding in short-lived species has population-level implications since the productivity in a population is determined in large part by the frequency of double brooding (Nagy and Holmes, 2004). At the individual level, a female's decision to invest in an additional reproductive attempt represents a trade-off between reproduction and her own survival plus any effects on the reproductive value of her first clutch (for example Nur, 1988; Verhulst, 1998; Bryant, 1979). Detection of such reproductive costs, however, is often obscured by confounding

^{*}Department of Mathematical Sciences/Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, 7491 Trondheim, Norway.

[†]Corresponding author. Email:jarle.tufto@ntnu.no

[‡]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Post Office Box 50, 6700AB Wageningen, Netherlands

effects such as territory quality and other environmental effects (Reznick, 1985; van Noordwijk and de Jong, 1986). Several studies have indeed reported higher survival rates of double-brooded females than single-brooded females (e.g. Boer-Hazewinkel et al., 1987) or survival being independent of the number of broods in a great tit population (Geupel and DeSante, 1990).

Given these trade-offs, females maximize their reproductive success by altering the number of offspring produced in each breeding attempt and the number of breeding attempts within each breeding season. Empirical studies have evaluated biotic and abiotic factors influencing multi-brooding strategies. For small passerine birds, there are probably a number of factors that influence variation among and within species in the number of broods laid each year including the sex ratio, territory quality, time of year, food supply, diet, body condition, age and experience. In a Dutch great tit (*Parus major*) population, Husby et al. (2009) showed that the frequency of double brooding declined over the study period, and that this decline is potentially related to the increasing phenological mismatch of each female's first clutch. For black-throated blue warblers, Nagy and Holmes (2005) finds that older females produce heavier nestlings and are more likely to double-brood. Another study of the same population of black-throated blue warblers found that early breeders were more likely to attempt a second brood than those starting later in a breeding season (Townsend et al., 2013).

Even though the observational studies of multi-brooding in different short-lived birds species have brought us better understanding of multi-brooding mechanisms, the findings from one species can not easily be extrapolated to other species. Additionally, different patterns may be found for different populations within the same species (e.g. Verhulst, 1998; Boer-Hazewinkel et al., 1987; Visser et al., 2003). Finally, some observational studies on natural populations seemingly do not agree with theoretical predictions. For example, in great tits, double brooding of great tits is phenotypically negatively correlated with laying date (of first brood). Given that laying dates have advanced by approximately 14 days (Charmantier et al., 2008), Husby et al. (2009) argues that this should have led to an increase in the frequency of double brooding, which is in marked contrast to the observed decline in the frequency of double brooding in the Hoge Veluwe population (Visser et al., 2003).

Despite the large body of empirical work on different factors affecting the frequency of multiple brooding, little theory on the evolution of multiple brooding exist. The aim of the present paper is therefore to formulate a simple quantitative genetic model for the evolution of double versus single brooding, assuming that the success of the first or both broods depends on their timing relative to a single window of favorable conditions reflecting for instance the abundance of a food resource. This makes the optimal onset of breeding (the laying date of the first brood) and the optimal number of broods to lay mutually dependent.

In a constant environment without frequency-dependent selection, mean population fitness as function of mean trait values can be viewed as an adaptive topography on which natural selection moves the population uphill (Lande, 1982). As we shall see, such a topography provides a useful tool for understanding the joint evolution of double brooding and onset of breeding. We first analyse under which conditions this topography has single or multiple adaptive peaks. Next, by deriving an adaptive topography under fluctuations in the resource peak (sensu Lande, 2007) we demonstrate that double brooding can arise as a conservative bet-hedging strategy in fluctuating environments, despite being selected against in a constant environment. To investigate possible effects of anthropogenic cli-

mate change on the frequency of double brooding, we also model the joint evolutionary trajectory under sustained movements of the resource window towards earlier dates. Finally, we discuss model predictions in light of parameter estimates obtained from the Hoge Veluwe long term data set on great tits.

2 Model and results

2.1 Assumptions

We assume that the relative contribution to fitness from a brood laid at time z is described by a Gaussian function $\exp\{-(z - \theta)/2\omega^2\}$ (as in Cao et al., 2019) reflecting a limited window in time of favorable breeding conditions, for example, the abundance of an important prey species (see for example Reed et al., 2013; Townsend et al., 2013). The parameter θ is the time at which conditions are the most favorable and ω specifies the width of the time window. We consider non-symmetric functions in the discussion.

We consider the joint evolution of two traits. The first trait z_1 is the date at which the first brood is laid (onset of breeding). Each female either lays one or two broods. We model this second trait as a threshold character (Falconer and Mackay, 1996, ch. 18) with an underlying liability z_2 . When the liability exceeds a threshold of zero, that is, for individuals for which $z_2 > 0$, a second brood is laid at time $z_1 + h$. If two broods are laid, fitness is reduced by a proportion c . With these assumptions, individual fitness expressed in terms of onset of breeding z_1 and the double-brooding liability z_2 is given by

$$w(z_1, z_2) = I_{z_2 \leq 0} e^{-\frac{(z_1-\theta)^2}{2\omega^2}} + I_{z_2 > 0} (1 - c) \left(e^{-\frac{(z_1-\theta)^2}{2\omega^2}} + e^{-\frac{(z_1+h-\theta)^2}{2\omega^2}} \right) \quad (1)$$

where I_A is an indicator function for the event A .

This model can be seen as a model with discrete but not necessarily non-overlapping generations as long as we assume age independent survival and fecundity rates and no density dependence. Our fitness measure $w(z_1, z_2)$ then corresponds to the multiplicative growth rate λ of phenotype z_1, z_2 . The Gaussian functions in (1) may be interpreted as the relative expected fecundities of all adult females (including offspring pre-census survival to age 1, i.e. whether the offspring recruit into the breeding population or not) and the cost c as the proportional reduction in both adult survival rates and fecundities. Assuming that the vectors of breeding and phenotypic values for the two traits, $\mathbf{x} = (x_1, x_2)^\top$ and $\mathbf{z} = (z_1, z_2)^\top$, are jointly multivariate normal, the change in mean phenotypic values from one generation to the next is given by

$$\Delta \bar{\mathbf{z}} = \mathbf{G} \nabla \ln \bar{w}(\bar{z}_1, \bar{z}_2) \quad (2)$$

where $\nabla = (\partial/\partial \bar{z}_1, \partial/\partial \bar{z}_2)^\top$ is the gradient operator and $\bar{w}(\bar{z}_1, \bar{z}_2)$ is the mean of (1) taken over the phenotype distribution of the population (Lande, 1982; Lande and Arnold, 1983; Caswell, 2001, ch. 11). The entries of the additive genetic variance-covariance matrix \mathbf{G} are defined in the usual way (Falconer and Mackay, 1996) as the variance and covariances of the additive genetic components of z_1 and the unobserved liability z_2 (instead of the binary threshold character itself). The evolution of \bar{z}_1, \bar{z}_2 can be understood as an upward climb on the adaptive topography defined by $\bar{w}(\bar{z}_1, \bar{z}_2)$. Making the usual assumption that \mathbf{G} remains approximately constant over time in addition to \mathbf{P} , evolutionary trajectories can be computed by iterating (2).

2.2 Adaptive topography for $P_{12} = 0$

Let P_{11}, P_{22} denote the phenotypic variances in z_1 and z_2 . For the moment, we also assume that the phenotypic covariance between z_1 and z_2 , $P_{12} = 0$. Factors involving z_1 and z_2 in (1) are then independent random variables and mean population fitness simplifies to

$$\bar{w}(\bar{z}_1, \bar{z}_2) = \sqrt{\frac{\omega^2}{\omega^2 + P_{11}}} \left(\Phi\left(-\frac{\bar{z}_2}{\sqrt{P_{22}}}\right) e^{-\frac{(\bar{z}_1-\theta)^2}{2(\omega^2+P_{11})}} + \Phi\left(\frac{\bar{z}_2}{\sqrt{P_{22}}}\right) (1-c) \left(e^{-\frac{(\bar{z}_1-\theta)^2}{2(P_{11}+\omega^2)}} + e^{-\frac{(\bar{z}_1+h-\theta)^2}{2(P_{11}+\omega^2)}}\right) \right), \quad (3)$$

where Φ is the standard normal cumulative density function (Appendix A). A contour plot of this function is depicted in Fig. 2 for four different parameter values.

Note first that (3) can be seen as a weighted average of mean fitness in populations consisting entirely of either single-brooding or double-brooding phenotypes. At intermediate \bar{z}_2 , mean fitness is thus bounded by mean fitness as function of \bar{z}_1 at $\bar{z}_2 = -\infty$ and $\bar{z}_2 = +\infty$ (the solid and dashed curve in Fig. A.1), respectively. Furthermore, note that $\bar{w}(\bar{z}_1, \bar{z}_2)$, for any given \bar{z}_1 , is either constant or a strictly monotonic function function of \bar{z}_2 . This implies that the adaptive topography never has any local or global peak anywhere else than at $\bar{z}_2 = \pm\infty$.

For the moment, we ignore complications of possible bifurcation of the double brood optimum happening for sufficiently large h (Appendix B). In a population consisting solely of double-brooding phenotypes ($\bar{z}_2 = +\infty$), setting $\partial\bar{w}(\bar{z}_1, +\infty)/\partial\bar{z}_1 = 0$, we then find that mean population fitness is then maximised when the mean timing of first and second broods are located symmetrically around the optimum θ at $\bar{z}_1 = \theta - h/2$ and $\bar{z}_1 + h = \theta + h/2$. This is a global adaptive peak if mean population fitness at these trait values is larger than maximum mean population fitness under single brooding, that is, when the inequality

$$\bar{w}(\theta - h/2, +\infty) > \bar{w}(\theta, -\infty), \quad (4a)$$

is satisfied. Applying (3) and rearranging, this translates into

$$\frac{h}{\sqrt{\omega^2 + P_{11}}} < 2\sqrt{2\ln(2-2c)}, \quad (4b)$$

depicted in Fig. 1, as the region left of the solid curve. Double brooding thus is a global optimum only if the time h between first and second broods is sufficiently small compared to the width of the resource peak corrected for the phenotypic variance in onset of breeding, $\omega^2 + P_{11}$. With an increasing cost c of laying two broods, the critical value of the delay h becomes smaller. If this cost reduce fitness by a proportion $c > 1/2$, double brooding is never a global optimum. Conversely, for parameter values for which (4b) is not satisfied, a single-brooding strategy with mean onset of breeding at the optimum, $\bar{z}_1 = \theta$, is the global adaptive peak.

For parameter values in the vicinity of the solid curve in Fig. 1, we shall see that the topography simultaneously may have adaptive peaks at both $\bar{z}_2 = -\infty$ and $\bar{z}_2 = +\infty$. Consider first the case where double-brooding is the global optimum such that (4a) holds. As illustrated by Fig. A.1, if

$$\bar{w}(\theta, +\infty) < \bar{w}(\theta, -\infty), \quad (5a)$$

it follows that mean population fitness $\bar{w}(\bar{z}_1, \bar{z}_2)$ is smaller than both $\bar{w}(\theta, -\infty)$ and $\bar{w}(\theta - h/2, +\infty)$ for all intermediate \bar{z}_1 in the interval $\theta - h/2 < \bar{z}_1 < \theta$. Hence, single-brooding

is an additional local peak. Conversely, if (5a) is not satisfied, the adaptive topography will always be strictly increasing towards the global maximum at $(\theta - h/2, +\infty)$ and no additional local peak can exist. Using (3), inequality (5a) translates to

$$\frac{h}{\sqrt{\omega^2 + P_{11}}} > \sqrt{2 \ln \frac{1-c}{c}}, \quad (5b)$$

the region to the right of the dotted curve in Fig. 1.

By a similar argument, if single-brooding is the global optimum, a second local double-brooding optimum exist whenever

$$\bar{w}(\theta - h/2, -\infty) < \bar{w}(\theta - h/2, +\infty), \quad (6a)$$

which leads to the condition

$$c < \frac{1}{2}, \quad (6b)$$

depicted by the region left of the dashed dotted line in Fig. 1. If the cost is too high ($c \geq 1/2$), there is no second local double-brooding optimum in addition to the global single brood optimum.

Fig. 2 shows four contour plots of the log of mean population fitness as function of \bar{z}_1 and \bar{z}_2 . As illustrated by Fig. 1, the four contour plots corresponds to the parameter values indicated by letters (a) to (d) in the parameter space. For parameter set (a) the cost is high ($c = 0.5$) and there is single and global, single-brooding peak at $\bar{z}_1 = \theta$. For parameter set (b) there are two peaks, one single-brooding and another double-brooding. For parameter set (c) there is a single global double-brooding peak.

Two evolutionary trajectories with assuming $\mathbf{G} = [\begin{smallmatrix} .4 & 0 \\ 0 & .4 \end{smallmatrix}]$ are shown in Fig. 2b, showing how the population may evolve towards different adaptive peaks depending on initial conditions. In practice, although a lengthy analytic expression for the gradient vector can be derived, we compute the the gradient vector by numerical differentiation in each generation.

2.3 Adaptive topography for $P_{12} < 0$

So far, in the derivation of (3) and in subsequent analysis, we have assumed that onset of breeding z_1 and the double-brooding liability z_2 are uncorrelated. Empirically, however, it is well known that the probability of double brooding tend to be higher for females that initiate breeding early (see e.g. Verboven and Verhulst, 1996; Brinkhof et al., 2002; Husby et al., 2009; Townsend et al., 2013). Such a negative phenotypic covariance may arise perhaps as a result of plasticity in both z_1 and z_2 or as a result of phenotypic canalization (as in Zhang and Hill, 2005) causing alignment of the phenotypic distribution with the individual fitness function (1).

In Appendix C, relying on the assumption that z_1, z_2 is jointly binormal as before, an exact expression for mean population fitness $\bar{w}(\bar{z}_1, \bar{z}_2)$ is derived similar to (3) but with the weights $\Phi(\bar{z}_2/\sqrt{P_{22}})$ replaced by more complex expression involving also \bar{z}_1 and h . Plots of this adaptive topography for the same parameter values as in Fig. 2 but with the phenotypic covariance changed to $P_{12} = -0.6$ is shown in Fig. 3.

Surprisingly, adaptive peaks at $\bar{z}_2 = \pm\infty$ are no longer present but have instead, except for parameter set (d), been replaced by a single global adaptive peak located at an intermediate mean liability \bar{z}_2 . In particular, for parameter set (b) and $P_{12} = 0$

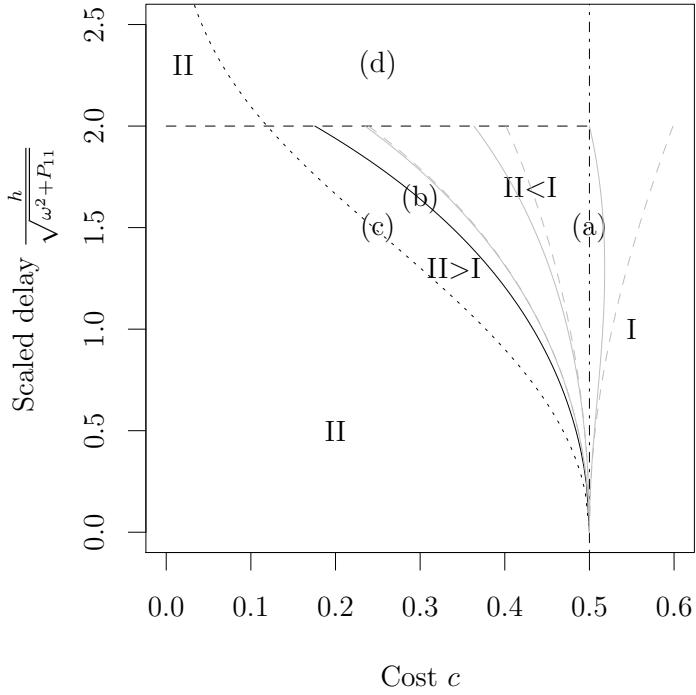


Figure 1: Model behaviour as function of the cost c of double brooding and the distance in time between the first and second brood h scaled against $\sqrt{\omega^2 + P_{11}}$ and for $P_{12} = 0$. Roman numerals indicate the existence of single- and double-brooding peaks in the adaptive landscape within different regions of the parameter space with inequality signs indicating global versus local peaks. The solid black curve thus delineate regions of the parameters space where single and double brooding are global adaptive peaks, respectively. The locations of letters a to d indicate the parameter values used in the corresponding subplots in Fig. 2. The grey curves relate to the model with fluctuations in the location θ of the resource window (Fig. 4). Like the solid black curve, they delineate regions of the parameter space where double and single brooding is the global optima for increasing values of $\sigma_\theta/\sqrt{\omega^2 + P_{11}} = 0.4, 0.8, 1.2$. The curves are either exact (solid grey curves computed numerically) or based on approximation (8) (dashed grey curves).

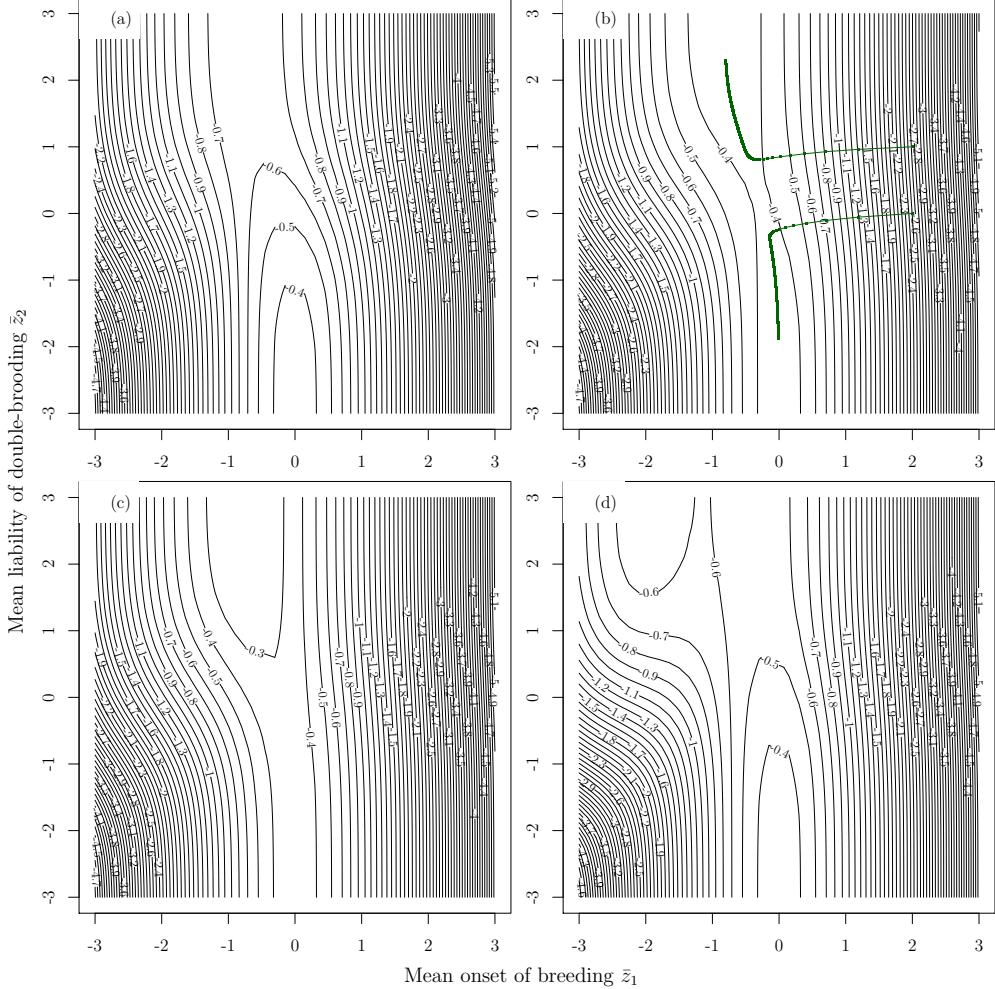


Figure 2: Adaptive topography for onset of breeding and liability of double brooding in a constant environment, $P_{12} = 0$. The contour lines are level curves of the log of mean population fitness as function of mean trait values. Each subplot (a)–(d) corresponds to parameter values shown in Fig. 1. The mean liability is shown in units of phenotypic standard deviations. Hence, $\bar{z}_2 = -2, -1, 0, 1, 2$ corresponds to proportions of double brooding equal to 0.02, 0.16, 0.5, 0.84, 0.97.

(Fig. 2b), disruptive selection for intermediate \bar{z}_2 causes the population to evolve towards either the adaptive peak at $\bar{z}_2 = -\infty$ or the one at $\bar{z}_2 = +\infty$. In contrast, for negative P_{12} (Fig. 3b) the two peaks merge into a single intermediate peak. For a single trait, it is well known that a bimodal individual fitness function with disruptive selection at intermediate trait values can translate into a single-peaked adaptive topography provided that the phenotypic variance is sufficiently large (Felsenstein, 1979). In the present model, the same phenomenon clearly occurs in the extreme case of an almost perfect phenotypic correlation such that phenotypes z_1, z_2 are distributed along a line intersecting the ridges emanating from each peak in Fig. 2b, and more generally for sufficiently strong phenotypic correlation between z_1 and z_2 .

The unimodal adaptive topography arising with sufficiently negative P_{12} thus implies that evolution can maintain a stable equilibrium population consisting of both single- and double-brooding phenotypes, as often seen empirically (Townsend et al., 2013; Husby et al., 2009). This occurs in the absence of any frequency-dependence. Of course, with sufficient gene flow, the maintenance of both phenotypes may also arise as a result of spatial variation in for example the width ω of the resource window or through spatial variation in other model parameters, such that single- and double-brooding constitute unimodal adaptive peaks at different spatial locations.

Note also that even single unimodal adaptive peaks located at $\bar{z}_2 = \pm\infty$ for $P_{12} = 0$ (Fig. 2a, c and d), to some extent are pulled toward intermediate \bar{z}_2 values (Fig. 3a, c and d) corresponding to proportions of double breeding of about 11, 67 and 6.6%, respectively.

2.4 Adaptive topography under fluctuating selection

Onset of breeding in birds is subject to fluctuating stabilizing selection (Chevin et al., 2015; Gamelon et al., 2018). In general, possible responses to such fluctuations include genetic evolution, plasticity and diversifying bet-hedging (see e.g. Tufto, 2015, and references therein). Here we show that double-brooding in part can be understood as another distinct mode of response to fluctuating selection. Keeping other assumptions unchanged, we assume that the location of the resource peak, instead of being constant, varies between years according to a normal distribution with mean μ_θ and variance σ_θ^2 . The geometric mean of mean population fitness across generations, or its log is then given by

$$\ln \bar{w}(\bar{z}_1, \bar{z}_2) = \int_{-\infty}^{\infty} \ln \bar{w}(\bar{z}_1, \bar{z}_2; \theta) f(\theta) d\theta \quad (7a)$$

where $f(\theta)$ is the probability density function of θ . Plots of (7a) (computed using numerical integration) are shown in Fig. 4, for the same parameter sets as in Fig. 3 and with $\sigma_\theta^2 = 1$ and $\mu_\theta = 0$.

The gradient of the adaptive topography defined by (7a) equals the expected selection gradient in any particular generation (Lande, 2007), since

$$\begin{aligned} \nabla \ln \bar{w}(\bar{z}_1, \bar{z}_2) &= \nabla \int_{-\infty}^{\infty} \ln \bar{w}(\bar{z}_1, \bar{z}_2; \theta) f(\theta) d\theta \\ &= \int_{-\infty}^{\infty} \nabla \ln \bar{w}(\bar{z}_1, \bar{z}_2; \theta) f(\theta) d\theta. \end{aligned} \quad (7b)$$

This gradient and the **G**-matrix thus determines the direction of expected evolution in any particular generation. Assuming that the expected evolution is realized every

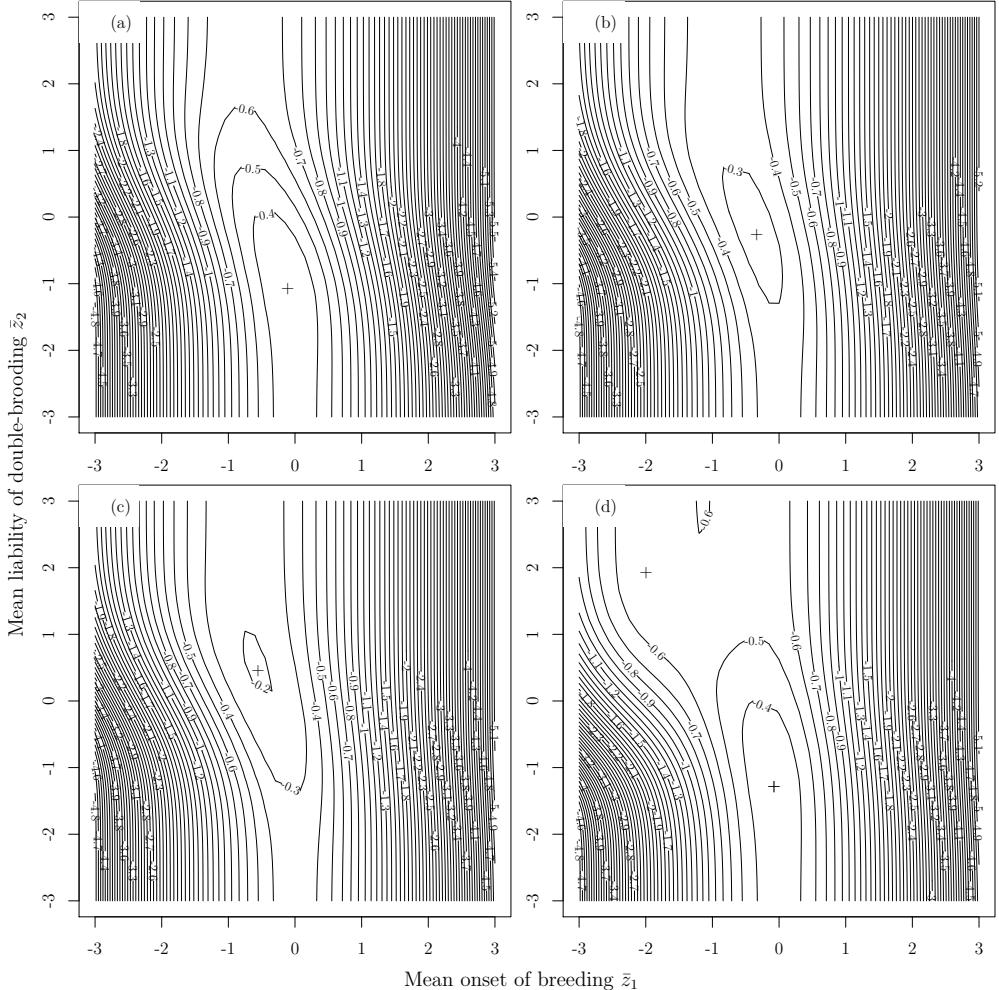


Figure 3: The same as Fig. 2 but with $P_{12} = -0.6$. The crosses indicate the location of adaptive peaks.

generation leads to the trajectories shown as blue curves in Fig. 4. The realized evolutionary trajectory (green curves), however, will be modified by stochastic variation in $\beta(\bar{z}_1, \bar{z}_2; \theta) = \nabla \ln \bar{w}(\bar{z}_1, \bar{z}_2; \theta)$. The expected evolutionary trajectories (the expectations of the green curves) are only approximately equal to the blue curves because the adaptive topographies are not entirely quadratic such that the gradient vectors are non-linear in \bar{z}_1, \bar{z}_2 .

Comparison with Fig. 3 reveals that fluctuations in θ in general shifts the locations of adaptive peaks toward higher frequencies of double brooding (larger \bar{z}_2). This additional adaptive benefit of double-brooding under fluctuating selection can essentially be understood as a conservative bet-hedging strategy. For example, consider the parameter values in Figs. 3d with a fixed resource peak at $\theta = 0$. Although maximum mean population fitness \bar{w} is higher under single brooding (having a log of about -0.35) than under double brooding (under which the log of maximum fitness is about -0.55), mean population fitness \bar{w} remain high under a wider range of \bar{z}_1 values under double compared to single brooding. With fluctuations in θ (shifting the horizontal location of $\bar{w}(\bar{z}_1, \bar{z}_2)$ in Fig. 3d), overall long term population fitness $\ln \bar{w}$ therefore become higher under double compared to single brooding (Fig. 4d), because mean population fitness in any particular generation becomes less variable when the location of the resource peak fluctuates. Essentially, laying multiple broods ensures that the timing of at least one brood is likely to match that of the resource peak.

Relying on the approximation in Appendix D and ignoring the complications of nonzero P_{12} , the condition for double brooding to be the global optimum changes from (4b) (solid black curve in Fig. 1) to the less restrictive approximate condition

$$\frac{h}{\sqrt{\omega^2 + P_{11}}} \lesssim 2 \sqrt{\frac{2 \ln(2 - 2c)}{1 - \sigma_\theta^2 / (\omega^2 + P_{11})}}, \quad (8)$$

depicted by the grey dashed curves in Fig. 1 for $\sigma_\theta / \sqrt{\omega^2 + P_{11}} = 0.4, 0.8, 1.2$. Comparison with the exact condition computed numerically (solid grey curves in Fig. 1) shows that approximation (8) works well unless σ_θ and h are both large (relative to $\sqrt{\omega^2 + P_{11}}$).

2.5 Linear trend in θ

During the last decades, as a consequence of warming spring temperatures, many bird species have advanced their onset of breeding (Crick et al., 1997) in response to selection for earlier laying dates (e.g. Gamelon et al., 2018). In great tits, the shift in optimal laying dates is mainly determined by a shift in the peak in caterpillar abundance (Visser et al., 2006; Chevin et al., 2015). These shifts in onset of breeding have only been partial, however, resulting in an increasing mismatch between the optimal and mean laying dates. Based on the negative phenotypic relationship between laying dates and probability of double brooding, Husby et al. (2009) hypothesize that we might expect an increase in the frequency of double brooding as laying dates advance to earlier dates. As we shall see, this is indeed a possible outcome if the G-matrix aligns with the phenotypic covariance matrix. Such alignment is expected from theory on phenotypic canalization and evolution of the G-matrix through mutation-selection balance (Arnold et al., 2008; Jones et al., 2014).

The effect of a linear deterministic trend in θ , possibly on top of stochastic fluctuations, is to increase the expected mismatch by a constant positive amount each generation. This will continue until this increase in the mismatch $\bar{z}_1 - \theta$ is balanced by the response to

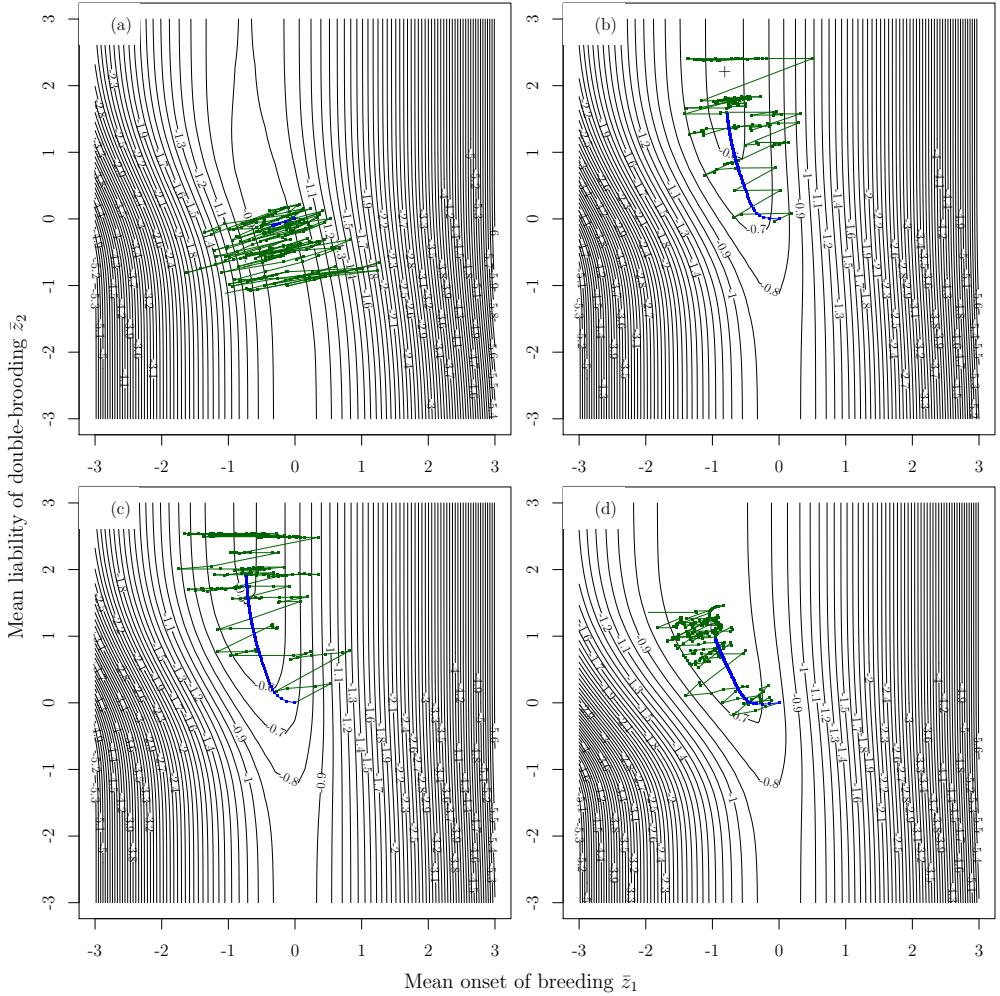


Figure 4: Log of geometric mean fitness for the same parameter sets as in Fig. 3 but with additional normally distributed fluctuation in the location θ . Also shown are 100 generations of realized and expected evolution (green and blue curves).

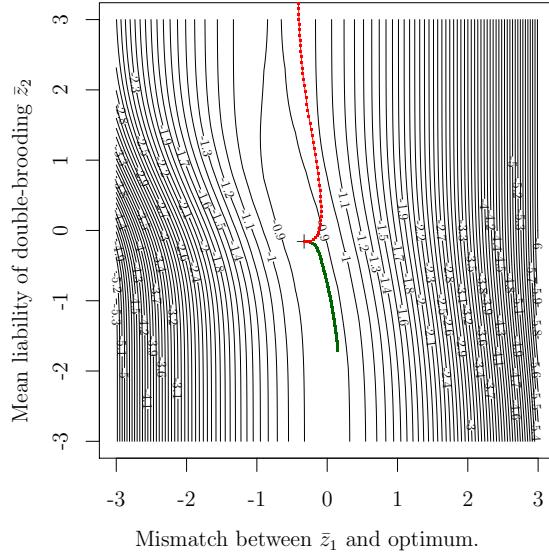


Figure 5: Expected evolution of phenotypic mismatch and mean liability of double brooding under sustained movements in θ towards earlier dates. Parameter values equal to those in Fig. 4a but with an additional deterministic shift in θ of -0.05 per generation. The green curve shows the evolutionary trajectory over 100 generations for zero genetic correlation between z_1 and z_2 and the red curve the trajectory for negative genetic correlation between z_1 and z_2 ($\mathbf{G} = 0.5\mathbf{P}$).

direct, directional selection in \bar{z}_1 in the opposite direction. If z_2 has a negative genetic correlation to z_1 , an indirect response to selection in \bar{z}_2 will therefore also occur in every generation. Hence, as long as the deterministic trend persists, an increase in double brooding should accumulate (Fig. 5, red curve). The evolving mean phenotype vector (\bar{z}_1, \bar{z}_2) thus rises above the horizontal line of the moving joint optimum, a phenomenon termed the ‘flying kite effect’ by Jones et al. (2004), also see Kopp and Matuszewski (2014). Unlike the evolving population considered by these authors, however, our ‘kite’ (the mean phenotype vector) do not necessarily settle at a constant ‘height’ since the adaptive topography becomes flat with respect to \bar{z}_2 for large \bar{z}_2 so the selection gradient may not have any downward \bar{z}_2 component preventing the ‘kite’ from rising indefinitely).

As the mean onset of breeding \bar{z}_1 begins to lag behind its optimum, however, this will disproportionately reduce the fitness contribution from any second brood, essentially because resource availability will be too low by the time a second brood can be initiated. This form of correlational selection creates direct selection for a reduced frequency of double brooding as can be seen from the downward orientation of the gradient vector to the right of the adaptive peak in Fig. 5. Hence, if there is no genetic correlation between z_1 and z_2 (or if the genetic correlation is positive), a reduced frequency of double brooding is expected to evolve in response to the trend in θ (green curve in Fig. 5). For moderate negative correlation between z_1 and z_2 , the resulting downward selection gradient may stabilize the ‘kite’ at a given ‘height’ \bar{z}_2 .

3 Empirical example

Using data from a long term study of the Hoge Veluwe great tit population, Cao et al. (2019) fitted a model similar to the one presented here, estimating the width of the resource window to $\hat{\omega} = 47.39$ days, the mean location θ of the resource window to $\hat{\mu}_\theta = 18.22$ days after April 1 (that is, April 18) and the standard deviation of θ to $\hat{\sigma}_\theta = 21.18$ days. The average time between successful first and second broods was estimated to $\hat{h} = 44$ days.

Other parameters that can be estimated from the same data are the phenotypic covariance matrix \mathbf{P} and mean trait values across the study period. To estimate P_{12} we assume that all entries of \mathbf{P} remain constant across years but that the means $\bar{z}_1 = \bar{z}_{1t}$ and $\bar{z}_2 = \bar{z}_{2t}$ vary across years. P_{11} can then be estimated by the usual pooled sample variance \hat{P}_{11} and \bar{z}_{1t} by yearly sample means \hat{z}_{1t} . Without loss of generality we assume that z_2 has a unit phenotypic variance $P_{22} = 1$. It then remains to estimate P_{12} and yearly mean liabilities \bar{z}_{2t} . Under the threshold model, using formulas for the conditional mean and variance of the multivariate normal distribution, the conditional probability p of double brooding given z_1 in a particular year t is

$$p = P(z_2 > 0 | z_1) = \Phi\left(\frac{\bar{z}_{2t} + P_{12}(z_1 - \bar{z}_{1t})/P_{11}}{\sqrt{1 - P_{12}^2/P_{11}}}\right), \quad (9a)$$

where Φ is the standard normal cumulative density function. This is a generalized linear model on the form

$$\text{probit } p = \Phi^{-1}p = \alpha_t + \beta z_1, \quad (9b)$$

with t and z_1 included as categorical and numerical covariates, respectively, that can be fitted to data using standard methods (McCullagh and Nelder, 1989).

Equating α_t and β in (9b) to corresponding quantities in (9a) and solving for P_{12} and \hat{z}_{2t} , the phenotypic covariance between onset of breeding and the liability for double brooding can be estimated by

$$\hat{P}_{12} = \frac{\hat{\beta} \hat{P}_{11}}{\sqrt{1 + \hat{\beta}^2 \hat{P}_{11}}}, \quad (10a)$$

and mean liability in year t by

$$\hat{z}_{2t} = \hat{\alpha}_t \sqrt{1 - \hat{P}_{12}^2/\hat{P}_{11}} + \hat{z}_{1t} \hat{P}_{12}/\hat{P}_{11}. \quad (10b)$$

Using this method on the Hoge Veluwe data we obtain $\hat{P}_{11} = 27.68$ (a phenotypic standard deviation of onset of breeding of 5.26 days), $\hat{P}_{12} = -1.59$ (a phenotypic correlation between onset of breeding and double brooding liability of -0.302). Estimates of \bar{z}_{1t} and \bar{z}_{2t} are shown in Fig. 6.

The parameter c in the above model remains an unknown parameter. Given the intermediate frequency of double brooding declining from an initial value of about 0.5 towards a value of 0.2 during the last two decades, it seems likely that the cost c must have an intermediate value such that neither single- nor double-brooding is strongly favored by selection. Fig. 7 shows the adaptive topography for a cost of double brooding $c = 0.45$ in addition to other parameter estimates given above. While this seems like a large cost, the conservative bet-hedging benefit of double brooding makes mean population fitness under double brooding about the same under single brooding.

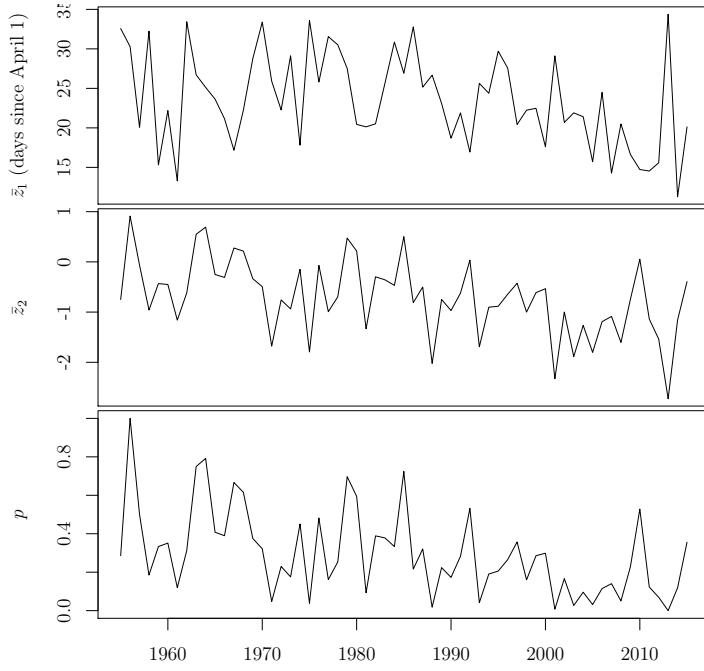


Figure 6: Estimated mean phenotypic values of onset of breeding \bar{z}_1 and liability of double brooding \bar{z}_2 and proportions of double brooding p across years in the Hoge Veluwe great tit population.

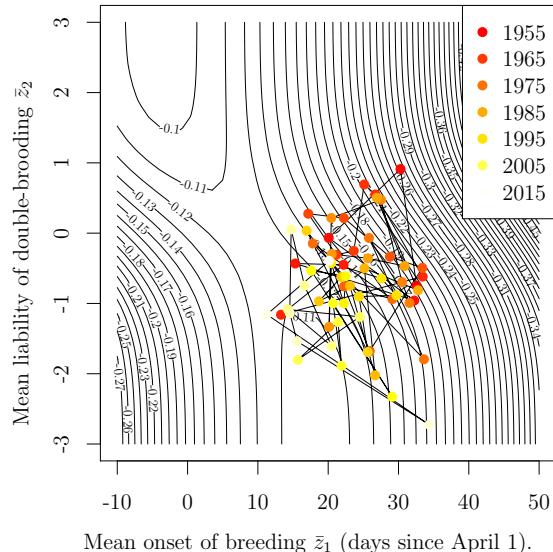


Figure 7: Adaptive topography for parameter values estimated for the Hoge Veluwe great tit populations assuming $c = 0.45$. The points show estimates of \bar{z}_1, \bar{z}_2 across different years.

Interestingly, even though we have estimated a negative phenotypic correlation between z_1 and z_2 , this correlation is too small to merge the two adaptive peaks into one. Instead the adaptive topography exhibit two peaks located at $\bar{z}_2 = \pm\infty$. The difference in mean population fitness at the two peaks and at the intermediate saddle point is small, however, only about 2%.

Over the course of the study period, the mean onset of breeding has lagged behind the optimal onset of by about 10 days. Relying on the value of $c = 0.45$ used in Fig. 7, the overall direction of evolutionary trajectory has also moved in the down- and leftward direction of the gradient of the mean topography. As discussed earlier, this downward direction of the selection gradient essentially arise because the fitness contribution from any second brood becomes diminished when the first brood is laid too late. Provided also that there is no strong negative genetic correlation between z_1 and z_2 , this thus provide a possible explanation for the observed decline in the frequency of double brooding in this population.

4 Discussion

Using a first principle-based quantitative genetic model describing the joint evolution of mean liability of double brooding and mean onset of double brooding under a scenario were successful breeding is limited to a finite and possibly fluctuating window in time, we have arrived at several novel insights. A simple but noteworthy prediction our model predict that the optimal onset of breeding relative to θ (for example a peak in prey abundance) should depend on the frequency of double brooding in the population, a complication usually ignored in empirical work (e.g Chevin et al., 2015; Ramakers et al., 2019) but see Cao et al. (2019). Comparison of laying dates of great tits with blue tits (*Cyanistes caeruleus*) (both depending on the peak in caterpillar abundance but the latter species rarely laying two broods) shows that the difference in mean laying dates between the two species changes with the changing frequency of double brooding in great tits (M.E. Visser, unpublished data), consistent with predictions of the current model.

Other, less intuitive predictions include the possible existence of adaptive topographies with multiple adaptive peaks, which in theory ultimately would maintain only a single strategy of pure single- or double-brooding. Alternatively, despite the absence of any frequency-depedence in our model, the topography may have a single intermediate peak maintaining a stable polyphenic population consisting of both single- and double-brooding individuals. Under fluctuations in θ , double brooding provides an additional conservative bet-hedging advantage. Finally, our model shows how the frequency of double brooding might change in either positive or negative direction under sustained trend in the location of the resource window. A reduced frequency of double brooding is expected if mean onset of breeding lags behind the optimum but this expectation can be reversed through a flying kite effect (sensu Jones et al., 2004) if the genetic correlation between onset of breeding and the liability of double brooding is sufficiently negative. A negative genetic correlation is expected if mutation causing earlier breeding also have pleiotropic positive effects on the probability of laying a second brood. Alignment of \mathbf{G} with $w(z_1, z_2)$ as expected from theory on mutation-selection balance (Arnold et al., 2008) would also cause G_{12} to evolve towards negative values.

We expect these model predictions to hold up at least qualitatively also more generally. We have made several simplifying model assumptions, however. First, we have not

included plasticity in onset of breeding z_1 . Importantly, there could also be plasticity in the liability z_2 of double brooding. Including plasticity in z_2 , one may speculate that double brooding to a larger extent may evolve to play the role of a backup mechanism, becoming expressed at a higher frequency in years where θ happen to fall on a late date. To model this would involve evolution in four reaction norm parameters. Restricting our model to non-overlapping generations, one approach would be to view the traits z_1 and z_2 as under direct selection (subject to the selection gradient computed numerically here) and with the reaction norm parameters (elevations and slopes) as under only indirect selection caused by genetic correlations with z_1 and z_2 , the technique employed by King and Hadfield (2019), p. A2. Environmental cues underlying plasticity in the liability of double brooding, however, must be different and only correlated with cues underlying plasticity in onset of breeding because more information about environmental conditions have accumulated at the time this decision is made. To some extent, although being ignored in empirical work on identification of climate windows (van de Pol et al., 2016), the amount of accumulated information must also vary at the individual level because of individual-level variation in z_1 and in h . This would further complicate formulas for \bar{w} and the realized selection gradient in any particular generation.

The current model has also ignored complications of failed broods. When such broods are followed by another breeding attempt, the separation in time is usually much smaller than the time interval h between successful broods (Cao et al., 2019) assumed in our simplified model. Failed broods would also incur different costs. In addition, some species (e.g. house sparrows) lay more than two broods (failed or successful) (Hegner and Wingfield, 1986) although this is an infrequent event in great tits considered in our empirical example. Both of these complications could be added to the present model but would complicate (1) and subsequent analysis. For example, a model with up to three broods would involve two thresholds in addition to the underlying liability, both of which could be evolving (in ways similar to the model of Chevin and Lande, 2013). We expect a long season of favourable conditions (large ω) as well as other mechanisms such as fluctuations in θ to favour more than two broods.

A final criticism that can be made is our use of a symmetric Gaussian function describing the window of favorable conditions. If this function is asymmetric with a steeper decline to the left, the optimal timing of two broods will no longer be centered symmetrically around the optimal laying date but becomes shifted to the right. In addition, with fluctuations in θ , an asymmetric function would further right-shift the mean onset of breeding through both insurance and conservative bet-hedging effects (Lof et al., 2012; Haaland et al., 2019). This could thus explain at least some of the mismatch between mean and optimal onset of breeding seen in Fig. 7.

Acknowledgements

We thank L.-M. Chevin for useful discussions and two anonymous reviewers for comments that greatly improved the manuscript. JT thanks CEFÉ-CNRS for hospitality during a sabbatical leave in 2018/19. YC was supported by the Research Council of Norway through its Center of Excellence funding scheme (project number 223257 to CBD)

References

- Arnold, S. J., R. Bürger, P. A. Hohenlohe, B. C. Ajie, and A. G. Jones, 2008. Understanding the evolution and stability of the G-matrix. *Evolution* 62:2451–61.
- Boer-Hazewinkel, D. et al., 1987. On the cost of reproduction: parent survival and production of second clutches in the great tit. *Ardea* 75:99–110.
- Brinkhof, M. W., A. J. Cavé, S. Daan, and A. C. Perdeck, 2002. Timing of current reproduction directly affects future reproductive output in European coots. *Evolution* 56:400–411.
- Bryant, D. M., 1979. Reproductive costs in the house martin (*Delichon urbica*). *Journal of Animal Ecology* 48:655–675.
- Cao, Y., M. E. Visser, and J. Tufto, 2019. A time series model for estimating temporal variation in phenotypic selection on laying dates in a Dutch great tit population. *Methods in Ecology and Evolution Accepted manuscript*.
- Caswell, H., 2001. Matrix Population Models. Sinauer, Sunderland, Massachusetts.
- Charmantier, A., R. H. McCleery, L. R. Cole, C. Perrins, L. E. B. Kruuk, and B. C. Sheldon, 2008. Adaptive phenotypic plasticity in response to climate change in a wild bird population. *Science* 320:800–803.
- Chevin, L.-M. and R. Lande, 2013. Evolution of discrete phenotypes from continuous norms of reaction. *American Naturalist* 182:13–27.
- Chevin, L.-M., M. E. Visser, and J. Tufto, 2015. Estimating the variation, autocorrelation' and environmental sensitivity of phenotypic selection. *Evolution* 69:2319–2332.
- Crick, H. Q. P., C. Dudley, D. E. Glue, and D. L. Thomson, 1997. UK birds are laying eggs earlier CO₂ increases oceanic primary production UK birds are laying eggs earlier. *Nature* 388:526.
- Falconer, D. and T. F. C. Mackay, 1996. Introduction to quantitative genetics. 4 ed. Longman, London.
- Felsenstein, J., 1979. Excursions along the interface between disruptive and stabilizing selection. *Genetics* 93:773–795.
- Gamelon, M., J. Tufto, A. L. Nilsson, K. Jerstad, O. W. Røstad, N. C. Stenseth, and B. E. Sæther, 2018. Environmental drivers of varying selective optima in a small passerine: A multivariate, multiepisodic approach. *Evolution* 72:2325–2342.
- Geupel, G. R. and D. F. DeSante, 1990. Incidence and Determinants of Double Brooding in Wrentits. *The Condor* 92:67–75.
- Haaland, T. R., J. Wright, J. Tufto, and I. I. Ratikainen, 2019. Short-term insurance versus long-term bet-hedging strategies as adaptations to variable environments. *Evolution* 73:145–157.

- Hegner, R. E. and J. C. Wingfield, 1986. Behavioral and endocrine correlates of multiple brooding in the semicolonial house sparrow *Passer domesticus* I. Males. *Hormones and Behavior* 20:294–312.
- Husby, A., L. E. B. Kruuk, and M. E. Visser, 2009. Decline in the frequency and benefits of multiple brooding in great tits as a consequence of a changing environment. *Proc. R. Soc. B* 276:1845–1854.
- Jones, A. G., S. J. Arnold, and R. Burger, 2004. Evolution and Stability of the G-Matrix on a Landscape with a Moving Optimum. *Evolution* 58:1639–1654.
- Jones, A. G., R. Bürger, and S. J. Arnold, 2014. Epistasis and natural selection shape the mutational architecture of complex traits. *Nature Communications* 5:3709.
- King, J. G. and J. D. Hadfield, 2019. The Evolution of Phenotypic Plasticity when Environments Fluctuate in Time and Space. *Evolution Letters* 3:15–27.
- Kopp, M. and S. Matuszewski, 2014. Rapid evolution of quantitative traits: Theoretical perspectives. *Evolutionary Applications* 7:169–191.
- Lande, R., 1982. A quantitative genetic theory of life history evolution. *Ecology* 63:607–615.
- , 2007. Expected relative fitness and the adaptive topography of fluctuating selection. *Evolution* 61:1835–1846.
- Lande, R. and S. J. Arnold, 1983. The measurement of selection on correlated characters. *Evolution* 37:1210–1226.
- Lof, M. E., T. E. Reed, J. M. McNamara, and M. E. Visser, 2012. Timing in a fluctuating environment: environmental variability and asymmetric fitness curves can lead to adaptively mismatched avian reproduction. *Proceedings of the Royal Society B* 279:3161–9.
- McCullagh, P. and J. A. Nelder, 1989. Generalized Linear Models, Second Edition.
- Nagy, L. R. and R. T. Holmes, 2004. Factors Influencing Fecundity in Migratory Songbirds : Is Nest Predation the Most Important? *Oikos* 35:487–491.
- , 2005. To Double-Brood or Not? Individual Variation in the Reproductive Effort in Black-Throated Blue Warblers (*Dendroica caerulescens*). *The Auk* 122:902.
- van Noordwijk, A. . J. . and G. . de Jong, 1986. Acquisition and Allocation of Resources : Their Influence on Variation in Life History Tactics. *American Naturalist* 128:137–142.
- Nur, N., 1988. The consequences of brood size for breeding blue tits. iii. measuring the cost of reproduction: survival, future fecundity, and differential dispersal. *Evolution* 42:351–362.
- van de Pol, M., L. D. Bailey, N. McLean, L. Rijsdijk, C. R. Lawson, and L. Brouwer, 2016. Identifying the best climatic predictors in ecology and evolution. *Methods in Ecology and Evolution* 7:1246–1257.

- Ramakers, J. J. C., P. Gienapp, and M. E. Visser, 2019. Phenological mismatch drives selection on elevation, but not on slope, of breeding time plasticity in a wild songbird. *Evolution* 73:175–187.
- Reed, T. E., S. Jenouvrier, and M. E. Visser, 2013. Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology* 82:131–144.
- Reznick, D., 1985. Costs of Reproduction: An Evaluation of the Empirical Evidence. *Oikos* 44:257.
- Townsend, A. K., T. S. Sillett, N. K. Lany, S. A. Kaiser, N. L. Rodenhouse, M. S. Webster, and R. T. Holmes, 2013. Warm springs, early lay dates, and double brooding in a north american migratory songbird, the black-throated blue warbler. *PLoS One* 8:e59467.
- Tufto, J., 2015. Genetic evolution, plasticity and bet-hedging as adaptive responses to temporally autocorrelated fluctuating selection: A quantitative genetic model. *Evolution* 69:2034–2049.
- Verboven, N. and S. Verhulst, 1996. Seasonal variation in the incidence of double broods: The date hypothesis fits better than the quality hypothesis. *Journal of Animal Ecology* 65:264–273.
- Verhulst, S., 1998. Multiple breeding in the great tit, ii. the costs of rearing a second clutch. *Functional ecology* 12:132–140.
- Visser, M. E., F. Adriaensen, J. H. Van Balen, J. Blondel, A. A. Dhondt, S. Van Dongen, C. Du Feu, E. V. Ivankina, A. B. Kerimov, J. De Laet, E. Matthysen, R. McCleery, M. Orell, and D. L. Thomson, 2003. Variable responses to large-scale climate change in European Parus populations. *Proceedings of the Royal Society B: Biological Sciences* 270:367–372.
- Visser, M. E., L. J. M. Holleman, and P. Gienapp, 2006. Shifts in caterpillar biomass phenology due to climate change and its impact on the breeding biology of an insectivorous bird. *Oecologia* 147:164–72.
- Weisstein, E. W., 2019. Convolution. From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/Convolution.html>.
- Zhang, X.-S. and W. G. Hill, 2005. Evolution of the environmental component of the phenotypic variance: stabilizing selection in changing environments and the cost of homogeneity. *Evolution* 59:1237–1244.

A Mean population fitness for $P_{12} = 0$

Using the fact that the expected value of a sum is the sum of the expected values and that expected values of products are products of expected values when z_1 and z_2 (and

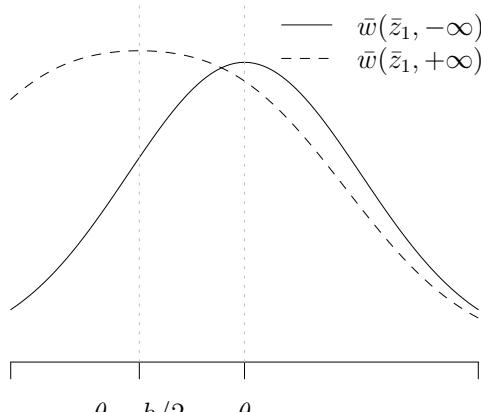


Figure A.1: Mean population fitness under pure single- or pure double-brooding (solid and dashed curve, respectively). For intermediate \bar{z}_2 , mean population fitness $\bar{w}(\bar{z}_1, \bar{z}_2)$ is a weighted average of the two curves and hence bounded between them. For the parameter values used, inequality (5a) holds such that $\bar{w}(\bar{z}_1, \bar{z}_2)$ has two adaptive peaks.

functions thereof) are independent, the expected value of (1) simplifies to

$$\begin{aligned}\bar{w}(\bar{z}_1, \bar{z}_2) &= Ew(z_1, z_2) \\ &= E \left(I_{z_2 \leq 0} e^{-\frac{(z_1-\theta)^2}{2\omega^2}} + I_{z_2 > 0} (1-c) \left(e^{-\frac{(z_1-\theta)^2}{2\omega^2}} + e^{-\frac{(z_1+h-\theta)^2}{2\omega^2}} \right) \right) \\ &= EI_{z_2 \leq 0} \cdot Ee^{-\frac{(z_1-\theta)^2}{2\omega^2}} + EI_{z_2 > 0} \cdot (1-c) \left(Ee^{-\frac{(z_1-\theta)^2}{2\omega^2}} + Ee^{-\frac{(z_1+h-\theta)^2}{2\omega^2}} \right)\end{aligned}\quad (\text{A.1})$$

Since z_2 is normally distributed with mean \bar{z}_2 and variance P_{22} , terms in (A.1) such as

$$EI_{z_2 > 0} = P(z_2 > 0) = P\left(\frac{z_2 - \bar{z}_2}{\sqrt{P_{22}}} > \frac{-\bar{z}_2}{\sqrt{P_{22}}}\right) = 1 - \Phi\left(-\frac{\bar{z}_2}{\sqrt{P_{22}}}\right) = \Phi\left(\frac{\bar{z}_2}{\sqrt{P_{22}}}\right).\quad (\text{A.2})$$

since $(z_2 - \bar{z}_2)/\sqrt{P_{22}}$ is standard normal.

Using well known formulas for Gaussian convolution integrals (Weisstein, 2019, eqs. 4-7), terms involving the expected value of Gaussian functions of normally distributed variables in (A.1) such as

$$Ee^{-\frac{(z_1-\theta)^2}{2\omega^2}} = \int_{-\infty}^{\infty} e^{-\frac{(z_1-\theta)^2}{2\omega^2}} \frac{1}{\sqrt{2\pi P_{11}}} e^{-\frac{(z_1-\bar{z}_1)^2}{2P_{11}}} dz_1 = \sqrt{\frac{\omega^2}{\omega^2 + P_{11}}} e^{-\frac{(\bar{z}_1-\theta)^2}{2(\omega^2 + P_{11})}}.\quad (\text{A.3})$$

Substituting (A.2), (A.3) and similar expressions into (A.1) leads to (3).

B Bifurcation of the double-brood optimum

If the delay h between the first and second brood is sufficiently large compared to the width of the resource window, it becomes optimal to initiate breeding such that either

the first or second brood is located close to the resource peak θ , that is, the double brood peak bifurcates into possible two adaptive peaks as illustrated by parameter set (d) in Fig. 2. The bifurcation occurs when the curvature of adaptive topography at $\bar{z}_1 = \theta - h/2$ and $\bar{z}_2 = \infty$ switches from stabilizing to disruptive selection, that is, when the second derivative $\partial^2 \bar{w} / \partial \bar{z}_1^2$ at this point turns positive. This happens when

$$\frac{h}{\sqrt{\omega^2 + P_{11}}} > 2, \quad (\text{B.1})$$

as depicted by the region above the dashed curve in Fig. 1. In a population with only double brooding, there are then two optima for \bar{z}_1 , one at a small distance $-\delta$ from θ and one at the same small distance δ from $\theta - h$. It follows that mean fitness at these peaks, $\bar{w}(\theta - \delta, \infty) = \bar{w}(\theta - h + \delta, \infty) \geq \bar{w}(\theta, \infty)$. Hence, a sufficient but not necessary condition for these optima to be the global is that $\bar{w}(\theta, \infty) > \bar{w}(\theta, -\infty)$, equivalent to the complements of conditions (5a) and (5b) depicted by the small region left of the dotted curve and above the dashed curve in Fig. 1. Essentially, if the delay h between the first and second brood is larger than (B.1), the cost have to be very small for double brooding to be a global peak of the adaptive topography.

C Mean fitness for non-zero P_{12}

For non-zero P_{12} , assuming that z_1, z_2 are jointly multivariate normal, z_1 conditional on z_2 is normally distributed with conditional mean

$$E(z_1|z_2) = \bar{z}_1 + P_{12}P_{22}^{-1}(z_2 - \bar{z}_2), \quad (\text{C.1})$$

and conditional variance

$$\text{Var}(z_1|z_2) = P_{11} - P_{12}^2 P_{22}^{-1}. \quad (\text{C.2})$$

Using (A.3) with the above mean and variance, the expectation of (1) conditional on z_2 is

$$\begin{aligned} E(w(z_1, z_2)|z_2) &= \sqrt{\frac{\omega^2}{\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1}}} \\ &\times \left[(1 - I_{z_2 \geq 0}) e^{-\frac{(\bar{z}_1 + P_{12}P_{22}^{-1}(z_2 - \bar{z}_2) - \theta)^2}{2(\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1})}} \right. \\ &\quad \left. + (1 - c) I_{z_2 \geq 0} \left(e^{-\frac{(\bar{z}_1 + P_{12}P_{22}^{-1}(z_2 - \bar{z}_2) - \theta)^2}{2(\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1})}} + e^{-\frac{(\bar{z}_1 + h + P_{12}P_{22}^{-1}(z_2 - \bar{z}_2) - \theta)^2}{2(\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1})}} \right) \right]. \end{aligned} \quad (\text{C.3})$$

Overall mean fitness can in turn be derived via the law of total expectation,

$$\bar{w}(\bar{z}_1, \bar{z}_2) = EE(w(z_1, z_2)|z_2). \quad (\text{C.4})$$

Substituting (C.3) into (C.4) yields an expression involving three integrals of Gaussian functions of z_2 over either $(-\infty, 0)$ or $(0, \infty)$. These can all be expressed in terms of the

standard normal cumulative density function. After some algebra, this leads to

$$\begin{aligned} \bar{w}(\bar{z}_1, \bar{z}_2) = & \sqrt{\frac{\omega^2}{\omega^2 + P_{11}}} \left[\Phi \left(- \frac{(\omega^2 + P_{11})\bar{z}_2 - P_{12}(\bar{z}_1 - \theta)}{\sqrt{P_{22}(\omega^2 + P_{11})(\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1})}} \right) e^{-\frac{(\bar{z}_1 - \theta)^2}{2(\omega^2 + P_{11})}} \right. \\ & + (1 - c)\Phi \left(+ \frac{(\omega^2 + P_{11})\bar{z}_2 - P_{12}(\bar{z}_1 - \theta)}{\sqrt{P_{22}(\omega^2 + P_{11})(\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1})}} \right) e^{-\frac{(\bar{z}_1 - \theta)^2}{2(\omega^2 + P_{11})}} \\ & \left. + (1 - c)\Phi \left(+ \frac{(\omega^2 + P_{11})\bar{z}_2 - P_{12}(\bar{z}_1 + h - \theta)}{\sqrt{P_{22}(\omega^2 + P_{11})(\omega^2 + P_{11} - P_{12}^2 P_{22}^{-1})}} \right) e^{-\frac{(\bar{z}_1 + h - \theta)^2}{2(\omega^2 + P_{11})}} \right]. \end{aligned} \quad (\text{C.5})$$

D Approximate benefit of double brooding under fluctuating selection

Under pure single-brooding, substituting (3) into (7a) and taking expectation we find that log of geometric mean fitness is

$$\overline{\ln \bar{w}}(\bar{z}_1, -\infty) = \frac{1}{2} \ln \left(\frac{\omega^2}{\omega^2 + P_{11}} \right) - \frac{(\bar{z}_1 - \mu_\theta)^2 + \sigma_\theta^2}{2(\omega^2 + P_{11})}. \quad (\text{D.1})$$

Similarly, under pure double-brooding, approximating $\ln \bar{w}(\bar{z}_1, +\infty)$ by a second order Taylor series in \bar{z}_1 around $\theta - h/2$, and taking expectation with respect to θ we find that

$$\begin{aligned} \overline{\ln \bar{w}}(\bar{z}_1, +\infty) \approx & \frac{1}{2} \ln \left(\frac{\omega^2}{\omega^2 + P_{11}} \right) + \ln 2 + \ln(1 - c) - \frac{h^2}{8(\omega^2 + P_{11})} \\ & - \frac{1}{2(\omega^2 + P_{11})} \left(1 - \frac{h^2}{4(\omega^2 + P_{11})} \right) \left((\bar{z}_1 + \frac{h}{2} - \mu_\theta)^2 + \sigma_\theta^2 \right). \end{aligned} \quad (\text{D.2})$$

Ignoring complications of bifurcation and nonzero P_{12} , double brooding is the global optimum if

$$\overline{\ln \bar{w}}(\mu_\theta - h/2, +\infty) > \overline{\ln \bar{w}}(\mu_\theta, -\infty). \quad (\text{D.3})$$

Substituting (D.1) and approximation (D.2) into (D.3) and solving for h leads to (8).

Multi-episodic fluctuating selection via fertility and viability in a great tit (*Parus major*) population

Yihan Cao¹, Marcel E. Visser², and Jarle Tufto¹

¹*Centre for Biodiversity Dynamics, Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway*

²*Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Post Office Box 50, 6700AB Wageningen, Netherlands*

CAO ET AL.— Estimating multi-episodic selection in a great tit population

Correspondence

Yihan Cao

E-mail: yihan.cao@ntnu.no

Author contributions

M.E.V. provided the data; J.T. conceived the idea; Y.C. initiated the statistical model, analyzed the data and conducted the analyses; Y.C. wrote the initial draft with input from J.T.; all authors contributed to revisions on later manuscript versions and gave final approval for publication.

Acknowledgments

We warmly thank Marlène Gamelon for her helpful and detailed comments on the initial draft. This work was supported by the Research Council of Norway through its Centres of Excellence funding scheme (project number 223257 to CBD).

Data Accessibility

All the necessary data and source code to carry out the analyses in this study will be available in a Dryad Digital Repository upon publication.

Abstract

Most previous studies working on varying selection over time either focused on single selective period, ignored temporal auto-correlation in selection, or did not account for ecological drivers of variation in selection. We formulated state-space models (SSMs) in which multiple selective episodes, potential variation and auto-correlation in selection as well as ecological variables driving selection are all integrated. We applied this framework to a long-term great tit population data with four episodes (offspring viability from egg to chick (episode 1), chick to fledgling (episode 2), fledgling to recruit (episode 3) and mother survival (episode 4) to estimate selection on laying date. The offspring viability for episodes 1 to 3 is decomposed into the multiplication of survival probability and non-zero inflation probability. We used Template Model Builder (TMB, R package) for model fitting and Akaike information criterion as a model selection criterion. Our selected model suggests that egg-laying date is under stabilizing selection through offspring survival with temporally auto-correlated and fluctuating optimal laying dates and directional selection through zero inflation probability. Beech crop size, clutch size, population density, food peak date all play a role driving selection in one or more episodes. Our multi-episodic approach of estimating selection can be directly applied to other study systems to uncover the whole picture of natural selection operating on the study species.

KEYWORDS: Multi-episodic selection, stabilizing selection, state-space model, Template Model Builder, zero-inflated beta-binomial regression.

1 Introduction

Given that fitness landscapes are not constant over time and populations are evolving towards a continuously changing fitness optimum (Fisher and Ford, 1947; Lande, 1976), empirical studies estimating fluctuating selection are accumulating rapidly (for example Chevin et al. (2015); Bonnet and Postma (2018) among many others). This also involves identifying ecological factors driving selection on different phenotypic traits over time and space. In spite of the importance of potential biotic and abiotic factors exerting selection, the empirical studies where we know the detailed ecological drivers of variation in selection are still scarce. Some notable exceptions include beak size in Darwin finches (Grant and Grant, 2002), reproductive timing in great tits (*Parus major*) (Husby et al., 2011), insecticide resistance in Culex mosquitoes (Milesi et al., 2016) and reproductive timing in Norwegian white-throated dippers (*Cinclus cinclus*) (Gamelon et al., 2018).

Even though the same (a)biotic covariates or autocorrelated environmental processes may act on multiple selective episodes, the mode, strength, and direction of selection on a trait may change from one life cycle segment to another (Arnold and Wade, 1984). Estimates of selection acting through one episode thus can not necessarily be extrapolated to another or even to lifetime selection. For example, Naef-Daenzer and Grüebler (2016) analyses weekly survival estimates from 123 data series based on studies of 65 species and suggests that using the number of fledglings as a proxy for fitness may be misleading in inference of overall

selection since reproductive success can be completely altered by many causal factors driving the adaptations, which operate during the post-fledgling period, and thereby change the juveniles' fate from fledgling to independence. Besides, it is intuitive to measure selection separately for episodes, for example, offspring viability, mating success, female fecundity, and survival. This might, nevertheless, only disclose part of the picture since the selection via these episodes are potentially correlated due to the shared environment the organism experiences over the life segments (Verhulst et al., 1995; de Forest and Gaston, 1996; Öberg et al., 2014). Therefore, the selection through different life stages should be estimated in a single statistical framework in which the correlation among selective episodes is taken into account by shared ecological variables and estimates of selection for different episodes are directly comparable, as done by Gamelon et al. (2018).

The pattern of temporal fluctuations in selection is characterized not only by its variance but also by its temporal auto-correlation (Tufto, 2015; Chevin and Haller, 2014; Chevin et al., 2015). Among the empirical studies accounting for temporal auto-correlation, Gamelon et al. (2018) proposed a multi-episodic approach where different reproductive stages (clutch size, offspring survival from egg to fledgling, from fledgling to recruit and breeding mothers) are included in one statistical model. Potential ecological drivers of selection on both laying dates and clutch size-laying dates were included and the method was applied to a dipper population.

Partial and complete brood failure are common in altricial birds (Santema and Kempenaers, 2018). Previous studies on nest productivity recognize multiple causes of nest failure include nest predation, nest abandonment that influence all the offspring (Etterson et al., 2007). Hatching failure and nestling mortality may be caused by infertility, by starvation or disease, which may affect only a subset of the offspring (Santema and Kempenaers, 2018). The proximate causes for variation in nest failure and mortality are often ignored but are important for understanding the natural selective processes and response of selection. Cao et al. (2019) proposed a statistical approach treating complete brood failure as a separate selective episode and showed that the model with this separation outperformed significantly the model without it. Therefore, it would be interesting to expand the model in Gamelon et al. (2018) to incorporate selective episodes via complete brood failure for both pre- and post-fledgling periods, such as chicks survival and recruitment probability, as well as a female's survival, into a comprehensive framework for estimating overall lifetime selection.

Take all of these together, we develop a statistical approach for estimating lifetime selection building on Gamelon et al. (2018) and Cao et al. (2019). In our study, the single episode in Cao et al. (2019) is extended to several episodes covering different life segments of a great tit and for each segment the complete brood loss is considered as a separate selective episode that is not done by Gamelon et al. (2018). By fitting the candidate time series models where the complete brood losses are or are not considered as separate selective episodes, together with a model fitting technique named Template Model Builder (TMB; Kristensen et al., 2016), we make the best use of the available data to explore a model formulation with lowest AIC value. We also show how ecological drivers such as spring temperature, breeding density, food resources abundance shape fluctuating selection which may operate differently via different reproductive cycle stages on the same phenotypic traits (egg-laying date and clutch size) of a great tit population.

2 Methods

2.1 Study species

The data we analyzed come from a great tit (*Parus major*) population at the Hoge Veluwe National Park in the Netherlands (52°02' - 52°07'N, 5°51' - 5°32'E). The study area consists of mixed pine-deciduous woodland on poor sandy soils and the great tit data has been collected since 1955 (see for example Reed et al., 2013; Townsend et al., 2013, for more details on the long-term study population). Given the raw data set, we excluded the records with manipulation, unknown mother and unknown brood type, and used a data set of $i = 1, 2, \dots, 5872$ breeding events from 1955 to 2015 in our statistical analysis. The number of eggs (y_{0i}), number of chicks (y_{1i}), number of fledglings (y_{2i}) and number of recruits (y_{3i}) were recorded for i th brood at its reproductive cycle stages and these fecundities are treated as fitness components that multiple selective episodes experienced. Mothers caught again breeding in the subsequent breeding season were considered as having survived and otherwise as dead (y_{4i} , 0-1 coded) so that mother local survival is considered as another fitness component.

We assume in this study that different selective episodes are similarly influenced by the same environmental covariates or random processes, even though the strength and the direction of selection on the egg-laying date may change from a reproductive cycle stage to another. Note that the number of eggs (i.e. clutch size) (y_{0i}) is considered in our study as a fitness component and focal trait that viability selection operates on simultaneously. In addition, we should have included age of mothers as a biological predictor in our study, however, 16.8% of the records in our date set employ unknown mother age and we thus decided not to model mother age for best utilization of the data.

2.2 Statistical model

The number of eggs (y_{0i}) is assumed to be Poisson distributed with parameter w_{0i} . We can see a high proportion of zeros in the number of chicks (6.56%), fledglings (15.50%) and recruits (74.91%) from Table 1 where the observed frequencies of the number of eggs, chicks, fledglings and recruits in our analyzed data are listed. We therefore used a zero-inflated beta-binomial (Zibb) model for each subsequent selective episode (y_{si} , $s = 1, 2, 3$) for offspring viability to capture the zero-inflated probability (zip_s) and over-dispersion in y_{1i} , y_{2i} and y_{3i} . The dispersion is parameterized by parameter γ_s for each episode, the survival probability is parameterized as p_{si} and the zero-inflation probability as $zip_{s'i}$ (to avoid confusion, in this study we refer p_{si} as offspring survival, $1 - p_{si}$ as offspring mortality, and $p_{si}(1 - zip_{s'i})$ as offspring viability). The joint distribution of y_{si} ($s = 0, 1, 2, 3$) is then

$$\begin{aligned} p(y_{0i}, y_{1i}, y_{2i}, y_{3i}) &= p(y_{0i})p(y_{1i}|y_{0i})p(y_{2i}|y_{0i}, y_{1i})p(y_{3i}|y_{0i}, y_{1i}, y_{2i}) \\ &= \text{Pois}(y_{0i}, w_{0i}) \prod_{s=1}^3 \text{Zibb}(y_{si}; y_{s-1,i}, zip_{s'i}, p_{si}, \gamma_s), \end{aligned} \quad (1)$$

where $\text{Zibb}(y_{si}; y_{s-1,i}, zip_{s'i}, p_{si}, \gamma_s)$ is the probability point mass function of zero-inflated beta-binomial distribution. In addition, the survival of the mother (y_{4i}) associated with

Table 1: Observed frequencies of number of eggs, chicks, fledglings and recruits in analysed great tit data set of this study.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Egg	32	7	16	38	92	197	440	758	1004	1157	1050	647	318	96	13	5	1	1	
Chick	385	45	98	178	244	442	694	876	940	900	581	341	119	24	4	1			
Fledgling	910	147	247	379	444	587	701	683	626	541	364	168	58	13	3	1			
Recruit	4399	996	327	92	39	15	4												

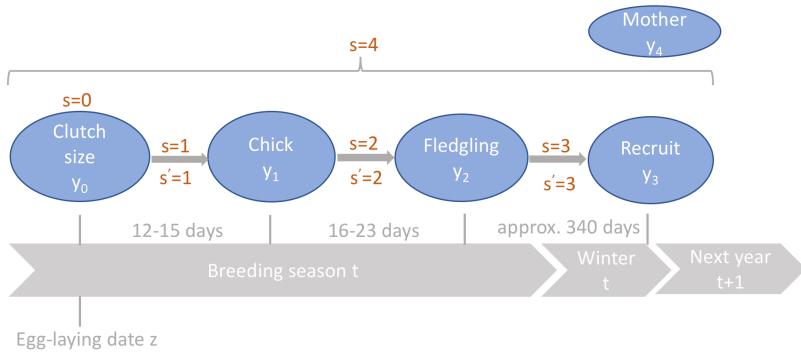


Figure 1: A diagram of multiple selective episodes considered in this study. $s = 0, 1, 2, 3, 4$ corresponds to five episodes from clutch size to mother survival. Specifically, $s = 0$ represents Poisson-distributed clutch size. The offspring viability from chick to recruit is decomposed into offspring survival probability ($s = 1, 2, 3$) and non-zero inflation probability ($s' = 1, 2, 3$) for each selective episode respectively. Selection through mother survival is denoted as episode $s = 4$. For a given brood, the clutch size, number of chicks, fledglings and recruits are y_0 , y_1 , y_2 and y_3 respectively. The mother survival is 0-1 coded as denoted as y_4 .

breeding event i is modelled with Bernoulli distribution with parameter p_{4i} (mother survival probability). As illustrated graphically in Figure 1, the selection via clutch size is considered as episode $s = 0$ and via mother survival as episode $s = 4$. The viability at each stage from egg to chick, chick to fledgling and fledgling to recruit is split into two episodes representing survival probability p_s and zero-inflation probability z_{ip_s} respectively. We can, therefore, see 8 selective episodes in total considered in this study.

We first model the parameter w_{0i} (the expected clutch size) and assume it to be under stabilizing selection of the form

$$w_{0i} = \exp \left(\eta_{0i}^{(\alpha)} - \frac{(z_i - \eta_{0i}^{(\theta)})^2}{2(e^{\eta_{0i}^{(\omega)}})^2} \right), \quad (2)$$

where $\eta_{0i}^{(\alpha)}$ is a predictor that might consist of fixed effects and random effects that determines

the maximum clutch size when the laying date z_i is at the optimal laying date $\eta_{0i}^{(\theta)}$. Parameter $\eta_{0i}^{(\omega)}$ determines the width of the stabilizing selection function, and thus the strength of selection.

Next, we model the survival probability p_{si} for $s = 1, 2, 3, 4$. As suggested by Ergon et al. (2018), a hazard rate $h(t)$ (t is age) is used to model p_{si} during episode s as

$$p_{si} = \exp \left(- \int_{t_{s-1}}^{t_s} h(t) dt \right), \quad (3)$$

where i indicates again the breeding event and t_{s-1} and t_s are the time of onset of episode $s-1$ and s respectively. The averaged mortality hazard rate at each selective episode s is then

$$\bar{h}_{si} = -\frac{\log(p_{si})}{t_s - t_{s-1}}. \quad (4)$$

A complementary log-log link is used to link the predictors and the survival probability p_{si} (the probability that a brood with (non-linear) predictor η_{si} will survive from episode $s-1$ to s), we thus get

$$\log p_{si} = -e^{\eta_{si}}. \quad (5)$$

The predictor η_{si} may contain fixed and random effects influencing the survival associated with the i th breeding event during selection episode $s = 1, 2, 3, 4$. Therefore, we assume that the model for cumulative hazard during each selective episode has the same form of stabilizing selection as on the laying date. Combining (4) and (5) we then have

$$\begin{aligned} \bar{h}_{si}(t_s - t_{s-1}) &= e^{\eta_{si}} \\ &= \exp \left(\eta_{si}^{(\alpha)} + \frac{(z_i - \eta_{si}^{(\theta)})^2}{2(e^{\eta_{si}^{(\omega)}})^2} \right), \end{aligned} \quad (6)$$

where $s = 1, 2, 3, 4$. Similar as (2), $\eta_{si}^{(\alpha)}$ is a parameter that determines the cumulative hazard in episode s when the laying date z_i is at the optimal laying date $\eta_{si}^{(\theta)}$. Parameter $\eta_{si}^{(\omega)}$ again determines the width of the stabilizing selection function. To investigate how the ecological variables drive multi-episodic selection in our great tit population, we link the potential biotic and abiotic factors with these three fitness parameters $(\eta_{si}^{(\alpha)}, \eta_{si}^{(\theta)}, \eta_{si}^{(\omega)})$ in (6) and the linear relationships are formulated as

$$\begin{aligned} \eta_{si}^{(\alpha)} &= \alpha_{0,s} + \alpha_{BCI2,s} BCI2_{t(i)} + \alpha_{BCI3,s} BCI3_{t(i)} + \alpha_{clusiz,s} y_{0i} + \alpha_{dens,s} \text{dens}_{t(i)} + \alpha_{mother,s} \xi_{j(i)}^{(\alpha)} + \alpha_{year,s} \mu_{t(i)}^{(\alpha)}; \\ \eta_{si}^{(\theta)} &= \theta_{0,s} + \theta_{temp,s} \text{temp}_{t(i)} + \theta_{peakd,s} \text{peakd}_{t(i)} + \theta_{dens,s} \text{dens}_{t(i)} + \theta_{clusize,s} y_{0i} + \theta_{year,s} \mu_{t(i)}^{(\theta)}; \\ \eta_{si}^{(\omega)} &= \omega_{0,s} + \omega_{temp,s} \text{temp}_{t(i)} + \omega_{year,s} \mu_{t(i)}^{(\omega)}, \end{aligned} \quad (7)$$

where t indicates breeding years in our analyzed dataset, $t = 1, 2, \dots, 61$. Parameters $\alpha_{0,s}$, $\theta_{0,s}$ and $\omega_{0,s}$ are episode dependent intercepts. Other fixed effects include beech crop index (3-point scale, dummy variables BCI2 and BCI3), clutch size (y_0), breeding density (dens), spring temperature (temp) and biomass peak date of caterpillars (peakd) (temp and peakd are highly correlated and only one of them should be kept in the selected model). More details on the fixed effects are given in Supporting Information.

In addition, we also add random mother effect ($\xi_{j(i)}$) on $\eta_{si}^{(\alpha)}$ and random temporal effect ($\mu_{t(i)}^{(\alpha)}, \mu_{t(i)}^{(\theta)}, \mu_{t(i)}^{(\omega)}$) on each of the fitness parameters. The subscript s for the coefficients in (7) indicates that the corresponding variable might affect the fitness parameters heterogeneously across the episodes. It is worth noting that the covariate clutch size is included in the linear model for $\eta_{si}^{(\alpha)}$ and $\eta_{si}^{(\theta)}$, meaning that the cumulative hazard and optimal laying date might be dependent on the clutch size and this translates into correlational selection on both laying date and clutch size. However, for episode $s = 0$ where the clutch size y_{0i} is considered as a fitness component, the covariate clutch size should be removed from the linear relationships in (7) ($\alpha_{\text{clusiz},0}$ and $\theta_{\text{clusiz},0}$ set to 0). Besides, for $s = 0$, we considered an alternative model where clutch size experienced directional selection on laying date with fixed intercept ($\alpha_{0,0}$), fixed slope ($\alpha_{\text{layd}}^{(1)}$), random temporal intercepts ($\alpha_t^{(0)}$, the standard deviation of which is denoted as $\sigma_\alpha^{(0)}$), and random temporal slopes ($\alpha_t^{(1)}$, the standard deviation of which is denoted as $\sigma_\alpha^{(1)}$). The correlation between random intercepts and slopes is denoted as ρ_α . The equation of directional selection is omitted to avoid complication.

Before going through assumptions for μ_t and ξ_j in (7), we formulate a linear model (the complete brood loss is experiencing directional selection) to link ecological factors to $\text{zip}_{s'i}$, the probability of extra zeros in y_{si} for episodes $s' = 1, 2, 3$ (we refer zero-inflated probability as "zip" for short in this paper). Similar as equation (3), (4) and (5), a complementary log-log link is used to link the zero-inflation probability and explanatory variables during episode s' :

$$\text{zip}_{s'i} = 1 - \exp(-\exp(\eta_{s'i}^{\text{zip}})), \quad (8)$$

and the linear model is given by:

$$\begin{aligned} \eta_{s'i}^{\text{zip}} = & \beta_{0,s'} + \beta_{\text{layd},s'} z_i + \beta_{\text{peakd},s'} \text{peakd}_{t(i)} + \beta_{\text{dens},s'} \text{dens}_{t(i)} \\ & + \beta_{\text{female},s'} \text{diffSurvFem}_{t(i)} + \beta_{\text{clusize},s'} y_{0i} + \beta_{\text{mother},s'} \xi_{j(i)}^{(\text{zip})} + \beta_{\text{year},s'} \mu_{t(i)}^{(\text{zip})}, \end{aligned} \quad (9)$$

where t indicates again breeding years and $\text{zip}_{0,s'}$ is the intercept in the model. Other fixed effects include breeding density (dens), clutch size (y_0) and annual survival difference between females who hatched and did not hatch chicks in year t is abbreviated as diffSurvFem (a proxy for mother disappearance, see Supporting Information for an explanation of the variables). Random mother effects $\xi_j^{(\text{zip})}$ and random year effects $\mu_t^{(\text{zip})}$ are again included to account for the variations between mothers and breeding seasons respectively. Alternative models with relative laying date (RLD=laying date+30-food peak date) included as an explanatory variable instead of brood laying date and food peak date were also tested. Note that subscript s' of the coefficients again indicates that the sizes of effects of the variables might differ across the episodes.

Furthermore, to model possibly auto-correlated and cross-correlated temporal fluctuations in the three fitness parameters, we assume that the random temporal effects $\mu_t^{(\alpha)}, \mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ are multivariate time series that can be modeled by an order 1 vector autoregression (VAR(1)) process, which is given by:

$$\begin{bmatrix} \mu_t^{(\alpha)} \\ \mu_t^{(\theta)} \\ \mu_t^{(\omega)} \end{bmatrix} = \Phi \begin{bmatrix} \mu_{t-1}^{(\alpha)} \\ \mu_{t-1}^{(\theta)} \\ \mu_{t-1}^{(\omega)} \end{bmatrix} + \mathbf{w}_t, \quad (10)$$

where \mathbf{w}_t (white noise) is assumed to be standard multivariate normal distribution $N(0, \Sigma)$ and Σ be an identity matrix, therefore $\mu_t^{(\alpha)}$, $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ reduce to three independent and identically distributed white noise processes when Φ turns out to be $\mathbf{0}$ matrix. The coefficients $\alpha_{\text{year},s}$, $\theta_{\text{year},s}$, and $\omega_{\text{year},s}$ in (7) allow the random temporal effects to vary across the episodes. Correlation between $\mu_t^{(\alpha)}$, $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ are determined through off-diagonal entries in Φ .

It is worth noting that, apart from the predictors regressed to $\eta_{si}^{(\alpha)}$ and $\eta_{s'i}^{zip}$, an offset term of the form $\log(\text{exposure})$, where exposure is the length of time for which a given brood is exposed to the possibility of failure, is added to $\eta_{si}^{(\alpha)}$ and $\eta_{s'i}^{zip}$ as an offset variable respectively.

2.3 Model selection and inference

The statistical model formulated in the last section (full model) is one of the models we implemented with the great tit data. TMB (Kristensen et al., 2016), an R package developed for fast fitting linear, non-linear and complex latent variables models is used for all the model fitting. We use Akaike's information criterion (AIC, Akaike, 1973) based on information theory to assess the data support for the candidate models. The best model with lowest AIC value is selected out through comparison among models with different degrees of complexities. The estimates of the parameters in the model, together with their standard errors were provided by the package.

To make sure the selected model is the best model among all potential models with different degree of complexity, the selected model (model S) was updated to neighbor models and the reported AIC values and the number of parameters kept in the neighbor models are compared with the selected model, as can be seen in Table 2. It is worth mentioning that we also fitted a model where the first episode $s = 0$ was under stabilizing selection and two other models where zero-inflation probabilities were disabled for episode $s = 2$ and $s = 3$ respectively, but the models eventually could not get converged due to the collapse of Laplace approximation used for integrating random effects out in TMB. This may happen when the random effects in (7) have to deal with the over-dispersion caused by zero-inflation. We thus no longer include them in Table 2. It is also worth noting in Table 2 that some models report very close AIC values to the best model retained but have fewer parameters (e.g. model N8, N11a, N19a). These models differ from the selected model that the statistically insignificant variables are removed from the selected model. This thus leads to less number of parameters in the neighbor models with the slightly worse model fit. Therefore, cautions need to be taken when interpreting the insignificant estimates reflected by standard errors of these parameters in the selected model.

3 Results and discussion

The best model indicates that there is no stabilizing selection operating on laying date for neither episode $s = 0$ (clutch size) nor episode $s = 4$ (mother survival). The alternative model with directional selection on laying date for $s = 0$ shows improvement in model fit (model N10 without directional selection worsen model fit), but the directional selection did not fluctuate over time (model N9 with random slopes indicates worse model fit). We

Table 2: Neighbour models of the selected model. Displayed are the tested models derived from the best model retained. The column ΔAIC and Δp list the difference in reported AIC value and the number of parameters between the selected model and each neighbor model respectively. The column Description gives the changes in each neighbor model has based on the selected model. Subscripts again indicate the selection episode under consideration, i.e. 1 indicate $s = 1$ and 1' represent $s' = 1$, so as to the other episodes.

Model	ΔAIC	Δp	Description
N1	2	2	$\theta_{0,0} = 0, \theta_{0,1} \neq \theta_{0,2} \neq 0, \theta_{0,3} = \theta_{0,4} \neq 0;$ $\omega_{0,0} = 0, \omega_{0,1} = \omega_{0,3} \neq 0, \omega_{0,2} \neq \omega_{0,4} \neq 0$
N2	4	3	$\theta_{0,0} = 0, \theta_{0,1} \neq \theta_{0,2} \neq 0, \theta_{0,3} = \theta_{0,4} \neq 0;$ $\omega_{0,0} = 0, \omega_{0,1} = \omega_{0,3} \neq 0, \omega_{0,2} \neq \omega_{0,4} \neq 0$
N3	1580.86	-5	disable zero-inflation probability for $s' = 1$
N4	6.84	4	$\alpha_{BCI2,0} \neq \alpha_{BCI2,1} \neq \alpha_{BCI2,2} \neq \alpha_{BCI2,3} \neq \alpha_{BCI2,4} \neq 0$
N4a	2.94	2	$\alpha_{BCI2,0} = \alpha_{BCI2,1} = 0, \alpha_{BCI2,2} \neq \alpha_{BCI2,3} \neq \alpha_{BCI2,4} \neq 0$
N4b	1.5	1	$\alpha_{BCI2,0} = \alpha_{BCI2,1} = \alpha_{BCI2,2} = 0, \alpha_{BCI2,3} \neq \alpha_{BCI2,4} \neq 0$
N5	2.09	3	$\alpha_{BCI3,0} \neq \alpha_{BCI3,1} \neq \alpha_{BCI3,2} \neq \alpha_{BCI3,3} \neq \alpha_{BCI3,4} \neq 0$
N6	1.85	1	$\alpha_{clusize,1} \neq \alpha_{clusize,2} \neq \alpha_{clusize,3} \neq \alpha_{clusize,4} \neq 0$
N7	2.89	3	$\alpha_{dens,0} \neq \alpha_{dens,1} \neq \alpha_{dens,2} \neq \alpha_{dens,3} \neq \alpha_{dens,4} \neq 0$
N8	0.67	-1	$\alpha_{mother,0} \neq \alpha_{mother,1} \neq \alpha_{mother,2} \neq 0, \alpha_{mother,3} = \alpha_{mother,4} \neq 0$
N9	117.33	2	add temporal random slopes of laying date on $\eta_0^{(\alpha)}$
N9a	251.31	-1	remove random temporal effects (random intercepts) from $\eta_0^{(\alpha)}$ for episode $s = 0$
N10	696.59	-1	remove $\alpha_{layd}^{(1)}$ from $\eta_0^{(\alpha)}$ for episode $s = 0$
N11	1.87	1	$\theta_{dens,0} = \theta_{dens,4} = 0, \theta_{dens,1} \neq \theta_{dens,2} \neq \theta_{dens,3} \neq 0$
N11a	0.78	-1	$\theta_{dens,0} = \theta_{dens,1} = \theta_{dens,2} = \theta_{dens,4} = 0, \theta_{dens,3} \neq 0$
N12	3.92	2	$\theta_{clusize,0} = \theta_{clusize,1} = \theta_{clusize,4} = 0, \theta_{clusize,2} \neq \theta_{clusize,3} \neq 0$
N12a	1.92	1	$\theta_{clusize,0} = \theta_{clusize,4} = 0, \theta_{clusize,1} = \theta_{clusize,2} = \theta_{clusize,3} \neq 0$
			remove peak date from $\eta_s^{(\theta)}$ for all the episodes $s = 0, 1, 2, 3, 4$;
N13	3.45	-9	$\theta_{temp,0} = \theta_{temp,4} = 0, \theta_{temp,1} \neq \theta_{temp,2} \neq \theta_{temp,3} \neq 0$
N14	1.45	1	$\omega_{0,0} = \omega_{0,4} = 0, \omega_{0,1} \neq \omega_{0,2} \neq \omega_{0,3} \neq 0$
N15	2.78	2	$\omega_{temp,0} = \omega_{temp,4} = 0, \omega_{temp,1} \neq \omega_{temp,2} \neq \omega_{temp,3} \neq 0$
N15a	1.93	0	$\omega_{temp,2} \neq 0, \omega_{temp,0} = \omega_{temp,1} = \omega_{temp,3} = \omega_{temp,4} = 0$
N16	23.08	-1	$\omega_{year,0} = \omega_{year,4} = 0, \omega_{year,1} = \omega_{year,2} = \omega_{year,3} \neq 0$
N16a	44.81	-2	remove random year effects from $\eta_s^{(\omega)}$ for all episodes $s = 1, 2, 3$
N17	1.98	1	$\beta_{0,1'} \neq \beta_{0,2'} \neq \beta_{0,3'} \neq 0$
N17a	7.61	-1	$\beta_{0,1'} = \beta_{0,2'} = \beta_{0,3'} \neq 0$
N18	0.96	1	$\beta_{layd,1'} \neq \beta_{layd,3'} \neq 0, \beta_{layd,2'} = 0$
N18a	2.95	2	$\beta_{layd,1'} \neq \beta_{layd,2'} \neq \beta_{layd,3'} \neq 0$
N19	0.83	1	$\beta_{clusize,1'} \neq \beta_{clusize,2'} \neq \beta_{clusize,3'} \neq 0$
N19a	0.25	-1	$\beta_{clusize,1'} \neq 0, \beta_{clusize,2'} = \beta_{clusize,3'} = 0$
N20	3.09	2	$\beta_{diffSurvFem,1'} \neq \beta_{diffSurvFem,2'} \neq \beta_{diffSurvFem,3'} \neq 0$
N21	1.58	3	$\beta_{dens,1'} \neq \beta_{dens,2'} \neq \beta_{dens,3'} \neq 0$
N22	2.69	2	$\beta_{mother,1'} \neq \beta_{mother,2'} \neq \beta_{mother,3'} \neq 0$
N23	49.4	-1	$\beta_{year,1'} \neq \beta_{year,3'} \neq 0, \beta_{year,2'} = 0$

thus conclude that clutch size is under directional selection on laying date and the selection gradient was constant across time, while we found no evidence that mother survival experienced phenotypic selection on laying date. The selected model also suggests that offspring viability from egg to recruit is zero-inflated and the survival probability experienced stabilizing selection while zero-inflation probability experienced directional selection, the selection strength for both fluctuated over the course of study. In this section, we discuss selection modes for each episode, effects of ecological variables, random effects on selection and explore the implications of these phenotypical selection processes operating on the focal traits. The estimates of the parameters in the selected model are shown in Table 3 and 4.

Due to the loglog-link between ecological factors and survival probabilities and the offset variable (log of time scale) added on $\eta_{si}^{(\alpha)}$ in (7), we get the time-averaged mortality hazard rate for each selective episode, which facilitates meaningful comparisons of ecological effect sizes across these selective episodes, irrespective of the episode length, as claimed in Ergon et al. (2018). For the same reason, the offset variable was added on (9) so that the effects of ecological variable on zero-inflated probabilities can also be compared straightforwardly across the reproductive cycle stages.

3.1 Ecological effects on clutch size and mother survival

We have mentioned that the best model indicates directional selection on laying date via $s = 0$ (clutch size) and the selection gradient was constant over the course of study. The estimate of $\hat{\alpha}_{\text{layd},0}^{(1)}$ to -0.00593 as can be seen in Table 3 suggests that one day earlier the mother started laying a brood, $0.0635 (e^{2.368+0.00593} - e^{2.368}) = 0.0635$ bigger the clutch size would be. The constant relationship between clutch size and laying date supports the 'absolute model' proposed by Winkler et al. (2002). The first row in Table 3 indicates that the maximum expected clutch size is equivalent to 10.69 ($e^{2.368} = 10.769$) with the other ecological variables being means. We find no evidence that the beech crop index level 2 (BCI2) nor level 3 (BCI3) have effects on clutch size, but breeding density shows a negative effect (-0.00126) on clutch size.

Since we find no evidence of phenotypic selection on reproductive timing via mother survival, the ecological variables only affect minimum hazard $\eta_{4i}^{(\alpha)}$ for episode $s = 4$. The estimate of $\hat{\alpha}_{0,4}$ to 0.278 translates the mother survival probability to $e^{-e^{0.278}} = 0.267$ in an average environment. Not surprisingly, both the beech crop levels have positive effects on mother survival. However, we find no evidence that clutch size and breeding density can affect adult female survival, implying no evidence of maternal cost of laying big clutches.

3.2 Ecological effects on offspring survival

We assumed that offspring survival experienced stabilizing selection on laying date since the clutches bred too early or too late would experience mismatch with food resource peak and thus suffer high mortality. Note again that offspring survival in this paper means p_s and viability means $p_s(1 - zip_s)$ for selective episode $s = 1, 2, 3$.

From Table 3 we can see that the offspring survival for $s = 3$ is markedly affected by beech crop and higher beech crop size leads to higher offspring survival from fledgling to recruit. The effect of BCI3 (-0.524) is more than two times larger than BCI2 (-0.233) for

Table 3: Estimates of the parameters in the linear predictor of fitness parameters for the clutch size ($s = 0$) and mortality (1-survival probability, $s = 1$ to 4) under the best model (model S in the model selection table in Supporting Information). Displayed are the notations of the parameters in (7), their estimates (\pm standard errors) and units where it is applicable for the selective episodes $s = 0, 1, 2, 3, 4$. The meanings of the parameters are listed as table notes. The estimates of parameters involved in VAR(1) process $\mu_t^{(\alpha)}$, $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ can be found in the text.

Parameter	Clutch size $s = 0$	Egg to chick $s = 1$	Selection episode Chick to fledgling $s = 2$	Fledgling to recruit $s = 3$	Adult female $s = 4$	Unit
$\hat{\alpha}_{0,s}$	2.368 ± 0.014	2.017 ± 0.086	0.143 ± 0.141	0.885 ± 0.076	0.278 ± 0.069	
$\hat{\alpha}_{BCI2,s}$	0	0	0	-0.233 ± 0.038	-0.233 ± 0.038	
$\hat{\alpha}_{BCI3,s}$	0	-0.128 ± 0.064	0	-0.524 ± 0.059	-0.249 ± 0.128	
$\hat{\alpha}_{clusize,s}$	0	-0.0810 ± 0.0094	0.0520 ± 0.0137	0.0159 ± 0.0063	0	
$\hat{\alpha}_{dens,s}$	-0.00126 ± 0.00032	0	-0.00215 ± 0.00079	0	0	
$\hat{\alpha}_{layd,s}^{(1)}$	-0.00593 ± 0.00026	0	0	0	0	day $^{-1}$
$\hat{\sigma}_{\alpha}^{(0)}$	0.0803 ± 0.00888	0	0	0	0	day $^{-1}$
$\hat{\alpha}_{mother,s}$	0.0107 ± 0.0059	0.2950 ± 0.0350	0.5804 ± 0.0380	0.0774 ± 0.0196	0.1337 ± 0.0293	
$\hat{\alpha}_{year,s}$	0	0.051 ± 0.019	0.194 ± 0.048	0.145 ± 0.033	0.295 ± 0.069	
$\hat{\theta}_{0,s}$	0	40.7 ± 2.6	33.5 ± 2.4	18.7 ± 3.1	0	day
$\hat{\theta}_{dens,s}$	0	-0.070 ± 0.042	-0.070 ± 0.042	-0.175 ± 0.053	0	
$\hat{\theta}_{peakd,s}$	0	1.36 ± 0.27	1.90 ± 0.21	2.45 ± 0.27	0	day $^{-1}$
$\hat{\theta}_{year,s}$	0	11.3 ± 2.8	11.3 ± 2.8	11.3 ± 2.8	0	
$\hat{\omega}_{0,s}$	0	3.878 ± 0.079	2.991 ± 0.037	3.878 ± 0.079	0	
$\hat{\omega}_{temp,s}$	0	0	0.058 ± 0.016	0.058 ± 0.016	0	
$\hat{\omega}_{year,s}$	0	0.416 ± 0.189	0.112 ± 0.053	0.112 ± 0.053	0	
$\hat{\gamma}_s$	0	1.638 ± 0.042	2.655 ± 0.106	1.023 ± 0.028	0	

- $\hat{\alpha}_{0,s}$: Intercept of $\eta_s^{(\alpha)}$
- $\hat{\alpha}_{BCI2,s}$: Effect of beech crop level 2 on $\eta_s^{(\alpha)}$
- $\hat{\alpha}_{BCI3,s}$: Effect of beech crop level 3 on $\eta_s^{(\alpha)}$
- $\hat{\alpha}_{clusize,s}$: Effect of clutch size on $\eta_s^{(\alpha)}$
- $\hat{\alpha}_{dens,s}$: Effect of density on $\eta_s^{(\alpha)}$
- $\hat{\alpha}_{layd,s}^{(1)}$: Fixed slope of laying date on clutch size ($s=0$)
- $\hat{\sigma}_{\alpha}^{(0)}$: Standard deviation of random intercepts of directional selection for $s=0$
- $\hat{\alpha}_{mother,s}$: Coefficient of random mother effects on $\eta_s^{(\alpha)}$
- $\hat{\alpha}_{year,s}$: Coefficient of random year effects on $\eta_s^{(\alpha)}$
- $\hat{\theta}_{0,s}$: Intercept of optimal laying date
- $\hat{\theta}_{dens,s}$: Effect of density on optimal laying date
- $\hat{\theta}_{peakd,s}$: Effect of caterpillar peak date on optimal laying date
- $\hat{\theta}_{year,s}$: Coefficient of random year effects on optimal laying date
- $\hat{\omega}_{0,s}$: Intercept of width of fitness function
- $\hat{\omega}_{temp,s}$: Effect of temperature on width of fitness function
- $\hat{\omega}_{year,s}$: Coefficient of random year effects on width of fitness function
- $\hat{\gamma}_s$: Over-dispersion parameter

episode $s = 3$. This finding validates the conclusion in Perdeck et al. (2000) studying the same population, which shows that both juvenile and adult survival is affected by BCI. Higher breeding density is related to smaller hazard mortality at stage chick to fledgling (estimate to -0.00215) but not at the other reproductive stages for offspring. The optimal laying date in an average environment is estimated to 40.7, 33.5 and 18.7 for $s = 1, 2, 3$ respectively. Breeding density is also associated with the optimal laying date negatively and the effects for the three episodes are -0.07, -0.07 and -0.175 respectively. Not surprisingly, the food peak date is positively related to optimal laying date for $s = 1, 2, 3$ and the estimates of the effect sizes are 1.36, 1.90 and 2.45 respectively. The width of fitness function in an average environment is estimated to 48.33 ($e^{3.878}$), 19.91 ($e^{2.991}$) and 48.33 ($e^{3.878}$) for episode $s = 1, 2, 3$ respectively, implying that the episode from chick to fledgling experienced the strongest selection. Spring temperature tends to affect the width of fitness function positively (estimate to 0.058) for episode $s = 2$ and 3 and this suggests that higher spring temperature leads to smaller strength of selection.

Different from Gamelon et al. (2018) where a quadratic effect of clutch size was present in (7), we did not find evidence of correlated selection operating on laying date and clutch size. Since we can see from Table 3 a comparatively strong negative effect of clutch size on offspring mortality from egg to chick (-0.081), positive effects from chick to fledgling (0.052) and from fledgling to recruit (0.0159), we speculate there is no optimal clutch size for a single selective episode but may exist optimal clutch size across the selective episodes $s = 1, 2, 3$, in the sense that the probability of a clutch surviving to produce any recruits ($p_1 p_2 p_3$) with assumption that it suffers no complete loss. To validate our speculation, we graphically show offspring survival against clutch size and laying late in an average environment (ecological variables set at mean values) in Figure 2. Specifically, the three plots in the left show the estimated survival probability for episode $s = 1, s = 2$ and $s = 3$ against egg-laying date and clutch size. The rightmost plot is the overall survival probability from $s = 1$ to 3 ($p_1 p_2 p_3$) against egg-laying date and clutch size. All the survival probabilities are computed with BCI1, within-year mean breeding density, mean temperature and mean food peak date with the best model and random year effects and random mother effects are integrated out. The first three plots validate our speculation that there exists no optimal clutch size for a single episode since the survival probabilities for these episodes are monotonous functions of clutch size. However, we find no optimal clutch size maximizing $p_1 p_2 p_3$ neither as can be seen from the fourth contour plot that the smaller the clutch size is, the higher the overall survival would be. The optimal laying dates shown from the first three plots agree with the estimates of $\theta_{0,s}$ in Table 3 and the overall estimated optimal laying date is sometime near the first of May in an average year.

Even though we did not find an intermediate clutch size that maximizes the overall offspring survival probability, it does not mean that the fecundities (the number of chicks, fledglings, and recruits) are not maximized for some intermediate clutch size, as expected through the trade-off between offspring number and offspring viability (Smith and Fretwell (1974), see also Gamelon et al. (2018)). Since in this study the offspring viability is decomposed into the multiplication of survival probability and non-zero inflation probability, the intermediate clutch size within the range of observed values may exist for maximized offspring viability for a single selective episode (e.g. $p_1(1 - zip_1)$) or over the three episodes ($p_1 p_2 p_3(1 - zip_1)(1 - zip_2)(1 - zip_3)$). However, the computation of viability values for a

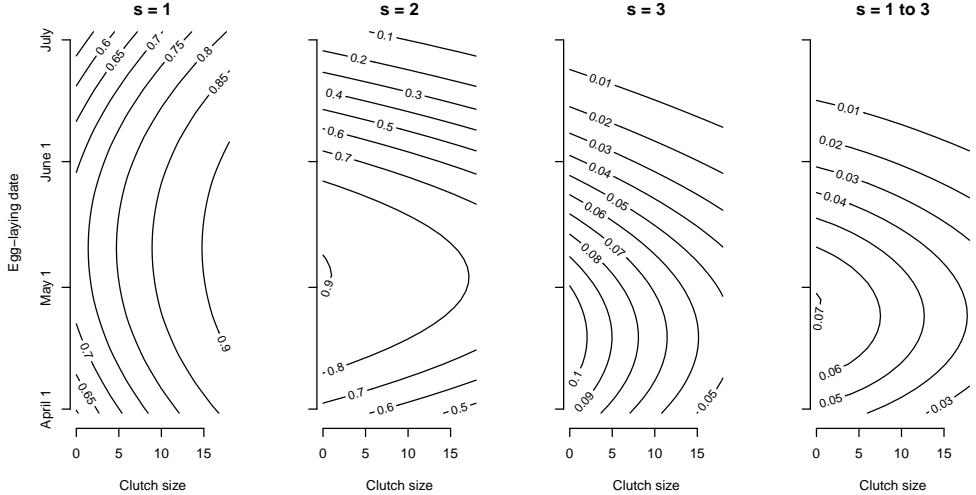


Figure 2: Contour plots of estimated offspring survival probability for episode $s = 1$ (from egg to chick), $s = 2$ (from chick to fledgling), $s = 3$ (from fledgling to recruit), and overall survival probability from $s = 1$ to $s = 3$ ($p_1 p_2 p_3$) against egg laying date and clutch size. The survival probabilities are computed with BCI1, with-in year mean breeding density, mean temperature and mean food peak date with the best model and random year effects, random mother effects are integrate out.

contour plot involves in integrating all the six random effects $(\mu_t^{(\alpha)}, \mu_t^{(\theta)}, \mu_t^{(\omega)}, \mu_t^{(zip)}, \xi_j, \xi_j^{(zip)})$ out and turns out to be extremely intensive, we are not able to show the contour plots similar to Figure 2 for offspring viability, instead, we calculate the estimated annual offspring viability with zero random mother effects, the details of which are given later.

3.3 Ecological effects on complete brood loss

From biological viewpoint, the excess of observed 0 in the offspring number should be regarded as another selection process independent of offspring mortality since the complete brood loss mainly results from predation for a lot of hole-nesting bird species (Bordjan and Tome, 2014; Maziarz et al., 2016; McCleery et al., 1996; Julliard et al., 1997; Naef-Daenzer et al., 2001; Wesołowski, 2002; Yamaguchi and Higuchi, 2005, for example) and offspring mortality from mismatch with food peak. Our analyzed data also support the model with zero-inflation probability taken as separate selective episodes, as can be seen from Table 2 that model N3 without zero-inflated probability for $s' = 1$ reports much worse model fit than our selected model. The models without zero-inflation probability for episode $s' = 2$ and 3 are even not converged as explained earlier.

From Table 4 we can see the estimates of parameters involved in zero-inflated probabilities for episodes $s' = 1, 2, 3$. Selection favors early broods for the episode from chick to fledgling and this is in line with the arguments from previous studies that juveniles from later broods suffer higher predation rates (Naef-Daenzer et al., 1999) and detrimental environmental

Table 4: Estimates of the parameters in the linear predictor for the zero-inflation probability under the best model (model S in the model selection table in Supporting Information). Displayed are the meaning of the parameters in (9), their notations, their estimates (\pm standard errors) and the units of parameters where it is applicable for selective episodes $s' = 1, 2, 3$. Note that these parameters stand for the effects on cloglog-transformed zip instead of on zip itself.

Meaning	Parameters	Symbol	Selection episode			Unit
			Egg to chick $s' = 1$	Chick to fledgling $s' = 2$	Fledgling to recruit $s' = 3$	
Intercept	$\hat{\beta}_{0,s'}$		-1.12 ± 0.13	0.19 ± 0.40	-1.12 ± 0.13	
Effect of difference in females survival	$\hat{\beta}_{\text{diffSurvFem},s'}$		-1.44 ± 0.42	0	0	
Effect of laying date	$\hat{\beta}_{\text{layd},s'}$		0.0117 ± 0.0034	0	0	day^{-1}
Effect of caterpillar peak date	$\hat{\beta}_{\text{peakd},s'}$		0	-0.047 ± 0.016	0	day^{-1}
Effect of clutch size	$\hat{\beta}_{\text{clusize},s'}$		-0.383 ± 0.038	-0.060 ± 0.040	0	
Standard deviation of random mother effects	$\hat{\beta}_{\text{mother},s'}$		5.48 ± 0.48	0	0	
Coefficient of random year effects	$\hat{\beta}_{\text{year},s'}$		0.23 ± 0.13	0.81 ± 0.12	0.40 ± 0.10	

conditions (Reed et al., 2013; Öberg et al., 2014). Surprisingly, clutch size shows negative effects on zip from egg to chick (estimate to -0.383) and from chick to fledgling (estimate to -0.060, not significant, but the candidate model without $\beta_{\text{clutchs},2'}$ (model N19a) report slightly worse model fit than the selected model), meaning bigger clutches suffered lower probability of complete brood loss for $s' = 1$ and 2 and the effect is much stronger for episode $s' = 1$. This finding might disagree what previous studies claimed that the larger/noisier broods are more detectable for predators (Perrins, 1965; Haskell, 1994; Leech and Leonard, 1997; Martin et al., 2000a; Götmark, 2002) and predators are attracted by more frequent parental visits to the nest (Martin et al., 2000b; Ferretti et al., 2005).

Similar to Figure 2, we use contour plots to show the non-zero inflation probability against laying date and clutch size, as can be seen in Figure 3 for each selective episode. For $s' = 1$ we can see an increase in the probability of survived broods (non-zero inflation probability) along with early laying dates and increased clutch size. While for $s' = 2$ only clutch size has an effect and for $s' = 3$ neither laying date nor clutch size has an effect on non-zip (thus no plot to show). These lead to an aggregating effect over $s' = 1, 2$ and 3 as shown on the rightmost plot that selection favors earlier and larger broods. This finding partly agrees with what found by Götmark (2002) studying on a Swedish great tit population that small broods suffered higher predation risk by sparrowhawks but they also showed that predation by sparrowhawks favors early breeding. Complete brood failure may also result from other reasons than predation for example bad weather leading to parental abandonment (Bordjan and Tome, 2014). Since we have no data for the timing of parents' disappearance, instead, we estimated the difference in survival rate between females that hatched and did not hatch chicks (diffSurvFem) and used it as a proxy of maternal disappearance, but we have no clue how close this estimation is to the reality. The records of the timing of parents' disappearance are required for a better understanding of the mechanism behind complete nest failure.

3.4 Random year effects and random mother effects

By turning on or off each entry of the transition matrix Φ , we could have more statistically reasonable candidate models with different Φ . To avoid a huge amount of potential candidate

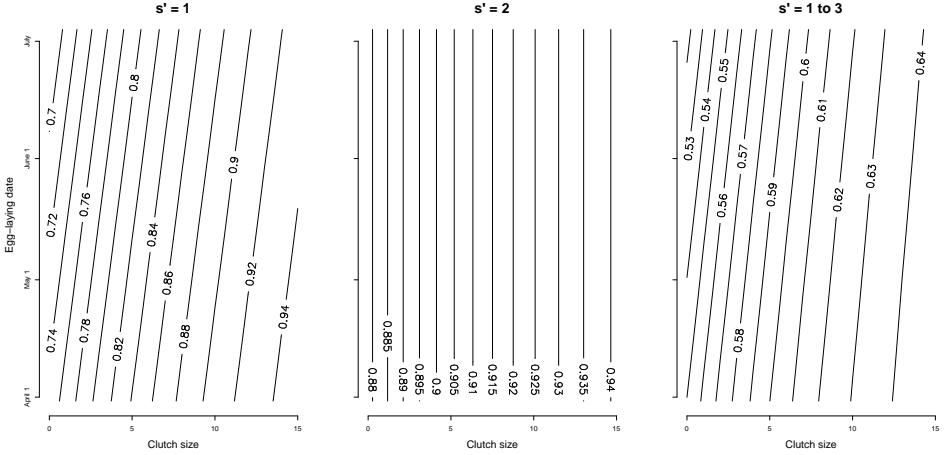


Figure 3: Estimated non-zero inflation probability on episode $s' = 1$ (from egg to chick), $s' = 2$ (from chick to fledgling), and overall non-zero inflation probability (multiplication of non-zip for $s' = 1, 2, 3$) against egg laying date and clutch size. The non-zip for $s' = 3$ is not shown since neither clutch size nor laying date suggests effect on non-zip. The non-zero inflation probabilities are computed at mean breeding density and mean difference of survival in females with our selected model and the random mother effects and random temporal effects are integrated out from the survival function and probability function.

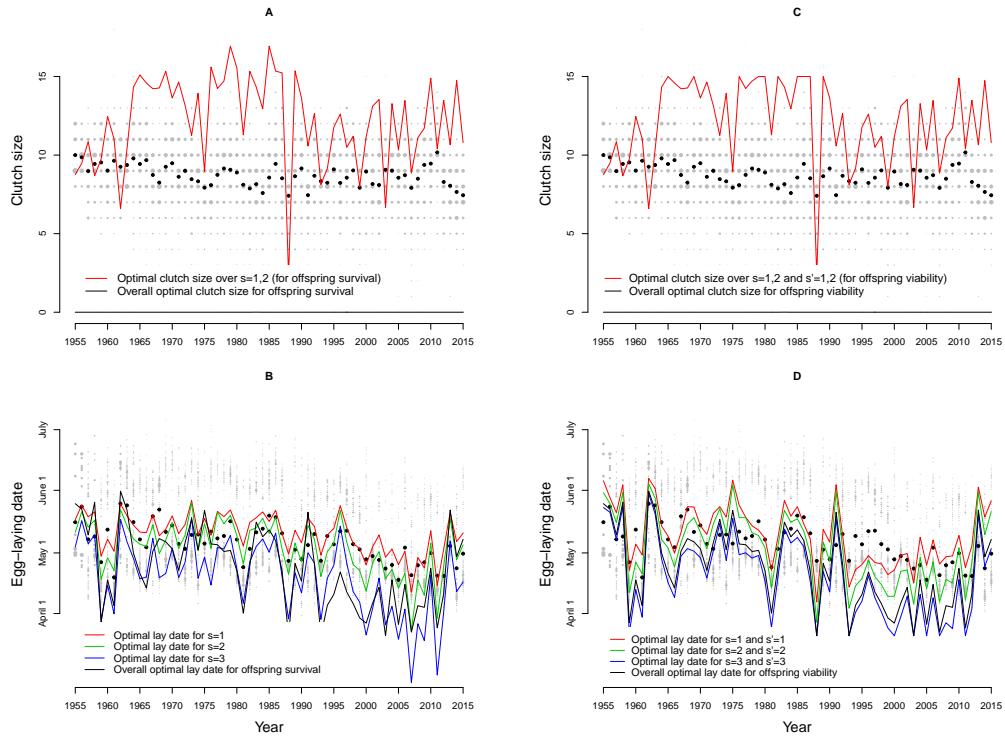
models, we instead estimate all the entries of Φ . The estimate (and standard deviation) of Φ in (10) with the selected model is $\hat{\Phi} = \begin{pmatrix} -0.29(0.21) & 0.14(0.20) & 0.84(0.66) \\ 0.30(0.19) & 0.49(0.13) & -0.23(0.20) \\ 0.19(0.27) & 0.15(0.16) & 0.64(0.18) \end{pmatrix}$ and the estimate (and standard deviation) of variance-covariance matrix of $\mu_t^{(\alpha)}$, $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ is $\begin{pmatrix} 1.38(0.33) & 0.32(0.36) & 0.13(0.67) \\ 0.32(0.36) & 1.61(0.69) & 0.59(0.76) \\ 0.13(0.67) & 0.59(0.76) & 3.18(2.34) \end{pmatrix}$. We find that only the estimates of the auto-correlation of $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ turn out to be significant (0.49(0.13) and 0.64(0.18) respectively). In addition, the variance-covariance matrix of $\mu_t^{(\alpha)}$, $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ turns to have only significant diagonal entries. Take these together, it is safe to claim that the process $\mu_t^{(\alpha)}$ is white noise while $\mu_t^{(\theta)}$ and $\mu_t^{(\omega)}$ are order-1 auto-regressive (AR(1)) process respectively. This conclusion agrees with the finding in Cao et al. (2019). The model N16a without random year effects on $\eta_s^{(\omega)}$ for episode $s = 1, 2, 3$ worsen the model fit and estimates of $\omega_{\text{year},s}$ are all significant for $s = 1, 2, 3$, we thus conclude that the width of fitness function fluctuated over the course of study for each selective episode from egg to recruit and there is more variation in random year effects for $s = 1$ (0.416) than episode $s = 2$ and 3 (0.112). The selected model also indicates significant random year effects on $\eta_{\alpha,s}$ for $s = 0, 1, 2, 3, 4$ as can be seen in Table 3. The shared random year effects for all the episodes implies that the year with larger clutches tends to have higher offspring mortality and breeding female mortality. Coefficient of random year

effects on the optimal laying date (η_θ) turned to be same for episodes $s = 1, 2, 3$ (estimate to 11.3). In Table 4, we can also see significant coefficients of random year effects in zero-inflated probabilities for all the three episodes $s' = 1, 2$ and 3 and there is more temporal variation for $s' = 2$ (0.81) than $s' = 1$ (0.23) and $s' = 3$ (0.40).

In Table 3 we can further see that the coefficients of random mother effects ($\alpha_{\text{mother},s}$) turn out to be significant for $s = 0, 1, 2, 3, 4$. Mothers show more variations in their offspring survival especially in the stage of egg to chick (0.295) and chick to fledgling (0.5804) than in the other selective episodes, which are 0.0107, 0.0774 and 0.1337 respectively for $s = 0, 3, 4$. It is worth noting that all the episodes share also the same random mother effects and this implies that the mother with bigger brood is assumed to suffer a higher offspring mortality and lower mother survival probability. Concerning zero-inflation probability, mothers behave differently to prevent their brood from complete loss only at the stage of egg to chick since only $\hat{\beta}_{\text{mother},1'}$ is significant.

3.5 Moving optimal laying date and clutch size

Figure 4 shows annual movement of optimal clutch size (plot A, C) and optimal egg-laying date (plot B, D) by maximizing offspring survival probability for each/several selective episode (p_1, p_2 and p_3). Optimal clutch size (plot C) and optimal egg-laying date (plot D) are calculated by maximizing viability for each/several selective episodes from egg to recruit ($p_1(1 - zip_1), p_2(1 - zip_2)$ and $p_3(1 - zip_3)$). Specifically, as we have found in Table 3 that the survival probability at $s = 1, 2$ and 3 is monotonous function of clutch size and this implies there is no optimal clutch size for a single selective episode, we thus in plot A shows optimal clutch size over episodes $s = 1, 2$ (solid red line) and over episodes $s = 1, 2, 3$ (solid black line). They were calculated by maximizing the survival probability (p_1p_2 over $s = 1, 2$ and $p_1p_2p_3$ for overall survival) as a function of clutch size and true ecological variables' values and estimates of all the effects (including estimated random year effects and zero random mother effects) were used. We also used the most frequently observed laying date 37 in the calculation. The same method applies to plot C. The grey dots are observations with size representing the frequencies of observations in the analyzed data and the black dots are the within-year means of clutch size. Surprisingly, we see from plot A large fluctuations in optimal clutch size (red line) ranging from around 3 to 16, which is even out of range of observations, for maximal survival probability across episode $s = 1$ and 2. Similar to Figure 2, there exists no optimal clutch size maximizing survival over $s = 1, 2, 3$ as we can see that the black line stays at 0 over the course of study. This implies that the negative effect of clutch size on survival probability for episode $s = 3$ markedly dominates the effects of clutch size on the survival for episode $s = 1$ and 2. We speculated that a confound effect of clutch size on survival probability and on non-zip would eventually lead to an intermediate optimal clutch size that maximizes the viability of a brood through reproductive stages from egg to recruit. However, we see from Plot C that the estimated optimal clutch size maximizing offspring viability over episode $s = 1$ and 2 (red line) shows similar moving pattern as in plot A, except for some "flat" part that implies the optimal clutch size is out of the space in which we search for optima. The optimal clutch size maximizing offspring viability for episode $s = 1, 2, 3$ (black) line also stays at 0 over the study years. This suggests that the negative effect of clutch size on survival for episode $s = 3$ not only dominates the effect of



(a) Maximized offspring survival probability.

(b) Maximized offspring viability.

Figure 4: Estimated optimal clutch size (A, C) and egg-laying date (B, D). In plot A and C, the grey dots represent observed clutch sizes over study years and the sizes of dots indicate the frequencies of the observations. The black dots are with-year mean of clutch sizes in plot A and C. The red and black lines represent estimated annual optimal clutch size accounting for episode $s = 1$ and 2 (from egg to fledgling) and episode $s = 1, 2, 3$ (from egg to recruits) respectively, with the selected model and egg-laying date=37, by maximizing survival in plot A and maximizing viability in plot C. In plot B, the grey dots indicate observed laying dates across the study period and the sizes again represent the frequencies of observations. The black dots are with-year mean laying dates. The red, green, blue and black lines represent estimated annual optimal laying date for episode $s = 1, 2, 3$ and overall (from egg to recruit) respectively, with the selected model and clutch size=9, by maximizing survival in plot B and viability in plot D. For all the plots, the random mother effects are set to 0 while the estimated random year effects and ecological effects are used.

clutch size on survival for $s = 1, 2$, it also dominates the effect of clutch size on zip for $s' = 1, 2$ and 3 .

Differently, there exists optimal laying dates for each reproductive stage $s = 1, 2$, and 3 , as shown in Plot B. The optimal laying date is obtained by maximizing survival probability (p_s for $s = 1, 2, 3$ and $p_1p_2p_3$ for overall survival) as a function of laying date where true ecological variables values, clutch size 9 (most frequent), estimated fixed effects and random year effects are used while the random mother effects are set to 0. The red, green, blue and black lines represent fluctuating optimal laying date for episode $s = 1, 2, 3$ and across three episodes over the course of study respectively. The observed laying dates are plotted with grey dots and the size indicates observed frequencies. It is interesting to see that the within-year mean of observed laying dates tracks closely either the red line or the green line, which means that adult females make brooding decisions using the environmental cues that only close to breeding time. Episode $s = 3$ suggests earlier laying dates as shown with the blue line and so as the overall optimal laying date. They are more fluctuating across years than the first two episodes but all of them showed a similar pattern and downward trend over the study years. We also find that the overall optimal laying dates are to a large degree determined by the optimal laying dates for $s = 3$, for most of the years. Similarly, the optimal laying dates in plot D are obtained by maximizing offspring viability $p_s(1 - zip_s)$ for $s = 1, 2$, and 3 and $p_1p_2p_3(1 - zip_1)(1 - zip_2)(1 - zip_3)$ for overall viability. The moving optimal laying dates show similar patterns as in plot B with relatively striking difference in the late years of study period, where the optimal laying dates for $s/s' = 3$ are less fluctuating and earlier in some years from 1995 to 2015, leading to bigger mismatch with yearly mean laying dates for these years.

3.6 Moving survival, zip, and recruit value

Since both survival and zip are affected by ecological variables and show randomness across study years, it is interesting to visualize the temporal trend of them and compare the trends across selective episodes. Figure 5 shows moving survival probability, non-zip probability and recruit (also chick, fledgling) values in plot A, B, C respectively. They are all calculated with the best model for a typical brood with clutch size 9 and laying date 37, the ecological variables and estimates of random year effects are used while random mother effects are integrated out. To be specific, plot A shows estimated annual survival probability. The red, green, blue and black lines represent survival for episode $s = 1, 2, 3$ and cumulative survival over these episodes respectively. We see that survival for $s = 1$ (red line) and 2 (green line) are close to each other and stay around 0.9 over years, while the survival for $s = 3$ (blue line) dropped dramatically below 0.2. We can also see the black line representing cumulative or overall survival across the three episodes almost overlaps the blue line, implying survival at $s = 3$ dominates $s = 1$ and 2 and contributes most to the overall survival.

Similar to plot A, plot B shows the estimated moving non-zip over study years for each selective episode and across episodes. What unexpected is that non-zip for $s' = 1$ (red line) is generally even smaller than $s' = 2$ (red line), both of which are fluctuating around 0.9. Non-zip for $s' = 3$ (blue line) again experienced more fluctuations than $s' = 1$ and 2 , dominated the other two episodes and stayed close to the cumulative non-zip (black line).

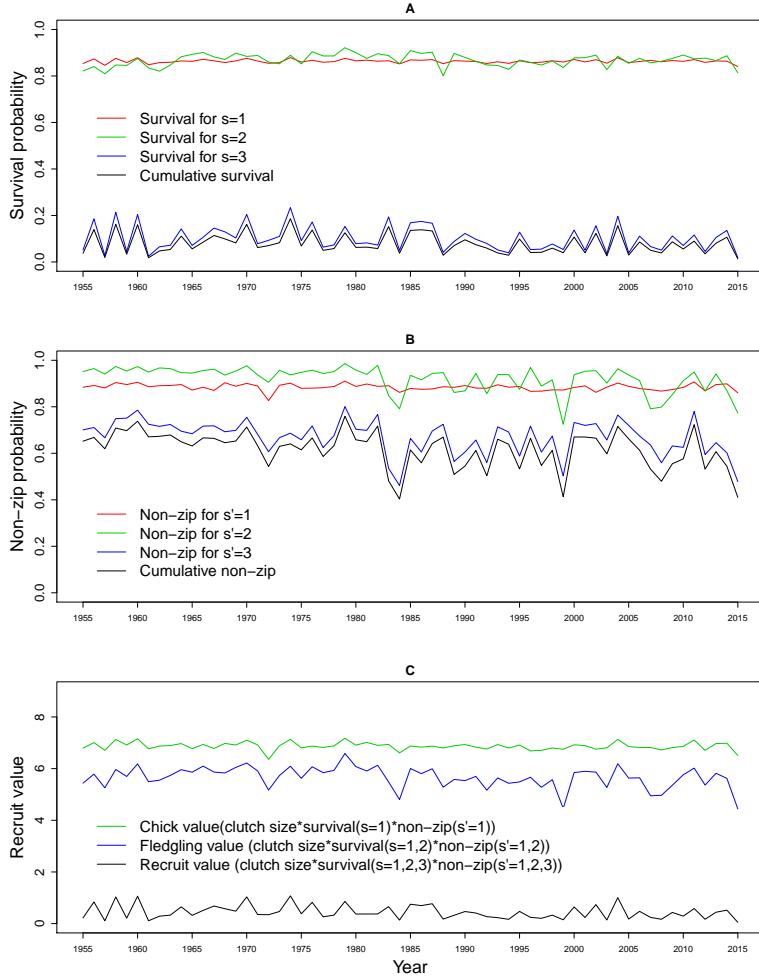


Figure 5: Estimated survival probability, non-zip, and recruit value for a typical brood (clutch size=9 and laying date=37) with the selected model. In plot A, the red, green, blue and black lines represent survival probability for episode $s = 1, 2, 3$ and cumulative survival probability across the three episodes (from egg to recruits) respectively. Similarly, plot B shows the non-zero inflation probability (non-zip) for episode $s' = 1, 2, 3$ and cumulative non-zip over the episodes with red, green, blue and black lines respectively. In plot C, the green, blue and black lines show the estimated number of chicks, fledglings, and recruits respectively for a typical brood. The chick values were calculated as multiplication of clutch size (=9), survival probability for $s = 1$ and non-zip for $s' = 1$. Similarly, the fledgling value is multiplication of clutch size (=9), cumulative survival for $s = 1$ and 2 and cumulative non-zip for $s' = 1, 2, 3$. The recruiting value is obtained by multiplying cumulative survival probability and non-zip over all the episodes from egg to recruit.

Plot A and B suggest that the excess 0 records of recruits result mainly from low survival from fledgling to recruit instead of complete brood failure.

At last, plot C shows estimated chick value (green line), fledgling value (blue line) and recruit value (black line) based on estimated survival and non-zip in plot A and B. The chick value is calculated as multiplication of a typical clutch size 9, p_1 and $1 - zip_1$. Similarly, fledgling value is obtained with $9 \cdot p_1 p_2 (1 - zip_1)(1 - zip_2)$, so on as to the recruit value. We find from the plot that for a typical brood with 9 eggs, approximate 7 survived to chicks and 6 survived to fledglings, but less than 1 survived to a recruit.

3.7 Model validation

The estimated survival probability and zip for each individual can be provided by TMB and this allows us to compare the observations with estimated values to evaluate the performance of our selected model. The five plots (A to E) in Figure 6 correspond to episode $s = 0$ to 4 respectively and plot B contains estimated values for both $s = 1$ and $s' = 1$, in plot C it is $s = 2$ and $s' = 2$ and in plot D $s = 3$ and $s' = 3$. The grey dots in the plots represent the within-year mean of observations and the meaning can be found in the legend of each plot. The size of the grey dots indicates the frequency of the observations. In plot A, the estimated clutch size (solid black line) with our selected model pass though almost all of the grey dots except for the first two years where only few data are available. In plot B, the red, green and black lines represent estimated within-year mean survival, non-zip, and viability (multiplication of survival and non-zip) respectively and the grey dots again indicate observed within-year mean viability for the episode from egg to chick. Our selected model estimates offspring viability very well for this episode since the black line goes through most of the grey dots. Similar explanation applies to plot C and D. We can see from plot D that the annual mean of viability are accurately estimated for $s = 3$ but for some years our model overestimated viability for $s = 2$ in plot C. Plot E shows estimated within-year mean mother survival with black line and for some years it goes though the observed within-year mean (grey dots). For the other years, there is a 0 to 0.2 discrepancy between the estimate and observation.

4 Conclusion

In this study, we extended the multi-episodic approach of estimating phenotypic selection from Gamelon et al. (2018) to include zero-inflation probabilities of offspring viability as separate selective episodes and claimed they have experienced different selective processes than offspring survival. We have fitted 87 candidate models and used AIC as a model selection criterion to select the best model. The best model suggests directional selection on laying date via complete brood failure and stabilizing selection operating on laying date via offspring survival at each reproductive stage from egg to chick, chick to fledgling, fledgling to recruit but not via mother survival. The optimal laying dates for each episode ($s = 1, 2, 3$) are determined by breeding density (negative effect) and food peak date (positive effect). The minimum offspring mortalities are negatively affected by beech crop levels. The width of the stabilizing selection function is positively affected by spring temperature only for $s = 2$

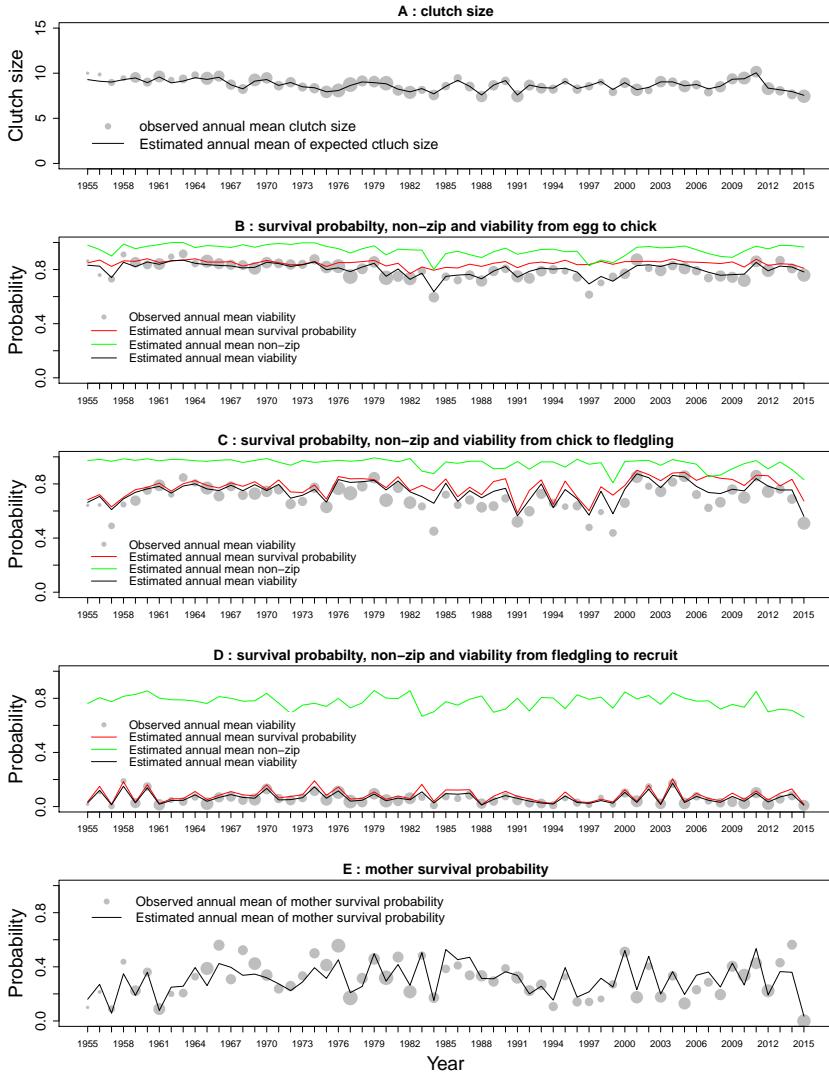


Figure 6: Observed data and estimated values with the selected model. Plot A shows the observed annual mean of clutch size in our analyzed data with grey dots and an annual mean of expected clutch size estimated with our selected model with a solid black line. Plot B shows observed annual mean viability from egg to chick with grey dots and the estimated annual mean of survival probability, non-zip and viability are shown with solid red, green and black lines respectively for the same reproductive stage. Similarly, plots C and D show these probabilities for reproductive stage chick to fledgling and from fledgling to recruit respectively. The observed and estimated annual mean of mother survival probability is shown in plot E with grey dots and a sold black line respectively. Note that the sizes of the grey dots in these plots all represent the frequency of the observations for the specific year and the frequencies are only comparable within the plot.

and 3. We also screened out biological and ecological factors affecting zip at each selective episode. The difference of survival between mothers hatched or did not hatch broods in the breeding season (a proxy for mother disappearance) shows a significant effect on zero-inflation probability for $s' = 1$. The early broods experienced lower complete brood loss risk also at $s' = 1$. Earlier caterpillar peak date is found related to higher zip at $s' = 2$ and the bigger brood suffered lower zip at the stage of $s' = 1$ and 2. Besides, the temporal fluctuations and autocorrelations are found in the optimal laying dates as well as in the width of the fitness function for selective episodes $s = 1, 2, 3$. Random mother effects turned to be significant for episodes $s = 0$ to 4 and $s' = 1$.

Besides, we provided another statistical framework for estimating selection on brood size using overall fitness involving both the parental and the offspring components as that in Tinbergen and Sanz (2004). McGlothlin (2010) proposed analytical methods to estimate lifetime nonlinear selection given quadratic and directional selection for each selective episode. We instead show that the life history selection can be calculated numerically by multiplying the selective episodes and integrating the random effects out. We have shown with contour plots that clutch size has negative effect on offspring survival over $s = 2$ and 3 but positive effect on non-zip over $s' = 1$ and 2, which potentially explains the contrasting findings from previous brood size manipulation studies that both enlarging and reducing brood size may lower fitness (Lindén, 1990; Tinbergen and Daan, 1990, for example). Furthermore, we find no evidence of correlated selection on the egg-laying date and clutch size via offspring survival and therefore find no intermediate clutch size maximizing offspring survival for every single episode but only over episode $s = 1$ and 2. Even though the cumulative offspring viability from egg to recruit tends to decrease with bigger clutch size, it does not mean that the selection favors small clutches since the number of actual survived recruits is a trade-off between clutch size and viability (see Supporting Information for a detailed discussion).

The low survival from fledgling to recruit (coupled with emigration but the information of which is not available) is the main reason for the low number of recruits. Fluctuating selection via offspring survival from fledgling to recruits also overwhelmingly dominates the selection via the other selective episodes. In our study, the probability surviving from fledgling to recruit is brood-based. It can be modeled at the individual level provided that required data are available since the recruit probability is dependent on individual-level variables such as fledgling mass (Monrós et al., 2002; Rodríguez et al., 2016). Even though we find no effects of reproductive timing and clutch size on mother survival directly, the significant variation of random mother effects on mother survival implies that the difference in mother survival ability is worth further investigation when data of biological variables of mother's quality is available.

References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *in* 2nd International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973, Pp. 267–281.
- Arnold, S. J. and M. J. Wade, 1984. On the measurement of natural and sexual selection: theory. *Evolution* 38:709–719.
- Bonnet, T. and E. Postma, 2018. Fluctuating selection and its (elusive) evolutionary consequences in a wild rodent population. *Journal of evolutionary biology* 31:572–586.
- Bordjan, D. and D. Tome, 2014. Rain may have more influence than temperature on nest abandonment in the great tit *parus major*. *Ardea* 102:79–87.
- Cao, Y., M. E. Visser, and J. Tufto, 2019. A time-series model for estimating temporal variation in phenotypic selection on laying dates in a dutch great tit population. *Methods in Ecology and Evolution* 10:1401–1411.
- Chevin, L.-M. and B. C. Haller, 2014. The temporal distribution of directional gradients under selection for an optimum. *Evolution* 68:3381–3394.
- Chevin, L.-M., M. E. Visser, and J. Tufto, 2015. Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. *Evolution* 69:2319–2332.
- Ergon, T., Ø. Borga, C. R. Nater, and Y. Vindenes, 2018. The utility of mortality hazard rates in population analyses. *Methods in Ecology and Evolution* 9:2046–2056.
- Etterson, M. A., L. R. Nagy, and T. R. Robinson, 2007. Partitioning risk among different causes of nest failure. *The Auk* 124:432–443.
- Ferretti, V., P. E. Llambías, and T. E. Martin, 2005. Life-history variation of a neotropical thrush challenges food limitation theory. *Proceedings of the Royal Society B: Biological Sciences* 272:769–773.
- Fisher, R. A. and E. B. Ford, 1947. The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity* 1:143–174.
- de Forest, L. N. and A. J. Gaston, 1996. The effect of age on timing of breeding and reproductive success in the thick-billed murre. *Ecology* 77:1501–1511.
- Gamelon, M., J. Tufto, A. L. Nilsson, K. Jerstad, O. W. Røstad, N. C. Stenseth, and B.-E. Sæther, 2018. Environmental drivers of varying selective optima in a small passerine: A multivariate, multiepisodic approach. *Evolution* 72:2325–2342.
- Götmark, F., 2002. Predation by sparrowhawks favours early breeding and small broods in great tits. *Oecologia* 130:25–32.
- Grant, P. R. and B. R. Grant, 2002. Unpredictable evolution in a 30-year study of Darwin's finches. *science* 296:707–711.
- Haskell, D., 1994. Experimental evidence that nestling begging behaviour incurs a cost due to nest predation. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 257:161–164.
- Husby, A., M. E. Visser, and L. E. Kruuk, 2011. Speeding up microevolution: the effects of increasing temperature on selection and genetic variance in a wild bird population. *PLoS biology* 9:e1000585.

- Julliard, R., R. H. McCleery, J. Clobert, and C. M. Perrins, 1997. Phenotypic adjustment of clutch size due to nest predation in the great tit. *Ecology* 78:394–404.
- Kristensen, K., A. Nielsen, C. W. Berg, H. J. Skaug, and B. Bell, 2016. TMB: automatic differentiation and Laplace approximation. *Journal of Statistical Software* 70:1–21.
- Lande, R., 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30:314–334.
- Leech, S. M. and M. L. Leonard, 1997. Begging and the risk of predation in nestling birds. *Behavioral Ecology* 8:644–646.
- Lindén, M., 1990. Reproductive investment and its fitness consequences in the great tit *Parus major*. Ph.D. thesis, Acta Universitatis Upsaliensis.
- Martin, T. E., P. Martin, C. Olson, B. Heidinger, and J. Fontaine, 2000a. Parental care and clutch sizes in north and south american birds. *Science* 287:1482–1485.
- Martin, T. E., J. Scott, and C. Menge, 2000b. Nest predation increases with parental activity: separating nest site and parental activity effects. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267:2287–2293.
- Maziarz, M., T. Wesołowski, G. Hebda, M. Cholewa, and R. K. Broughton, 2016. Breeding success of the Great Tit *Parus major* in relation to attributes of natural nest cavities in a primeval forest. *Journal of Ornithology* 157:343–354.
- McCleery, R., J. Clobert, R. Julliard, and C. Perrins, 1996. Nest predation and delayed cost of reproduction in the great tit. *Journal of Animal Ecology* Pp. 96–104.
- McGlothlin, J. W., 2010. Combining selective episodes to estimate lifetime nonlinear selection. *Evolution* 64:1377–1385.
- Milesi, P., T. Lenormand, C. Lagneau, M. Weill, and P. Labbé, 2016. Relating fitness to long-term environmental variations in natura. *Molecular ecology* 25:5483–5499.
- Monrós, J. S., E. J. Belda, and E. Barba, 2002. Post-fledgling survival of individual great tits: the effect of hatching date and fledgling mass. *Oikos* 99:481–488.
- Naef-Daenzer, B. and M. U. Gruebler, 2016. Post-fledgling survival of altricial birds: ecological determinants and adaptation. *Journal of Field Ornithology* 87:227–250.
- Naef-Daenzer, B., F. Widmer, and M. Nuber, 1999. Is the timing of breeding in great and coal tits adapted to maximum nestling food or to avoid juvenile predation. *in Proceedings of the 22 International Ornithological Congress*, Durban 1999, Pp. 1406–1415. Bird Life.
- , 2001. Differential post-fledgling survival of great and coal tits in relation to their condition and fledgling date. *Journal of Animal ecology* 70:730–738.
- Öberg, M., T. Pärt, D. Arlt, A. T. Laugen, and M. Low, 2014. Decomposing the seasonal fitness decline. *Oecologia* 174:139–150.
- Perdeck, A. C., M. E. Visser, and J. H. Van Balen, 2000. Great tit *parus major* survival and the beech-crop. *Ardea* 88:99–106.
- Perrins, C. M., 1965. Population fluctuations and clutch-size in the Great Tit, *Parus major* L. *The Journal of Animal Ecology* Pp. 601–647.
- Reed, T. E., S. Jenouvrier, and M. E. Visser, 2013. Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology* 82:131–144.

- Rodríguez, S., A. J. Van Noordwijk, E. Álvarez, and E. Barba, 2016. A recipe for postfledging survival in great tits *parus major*: be large and be early (but not too much). *Ecology and evolution* 6:4458–4467.
- Santema, P. and B. Kempenaers, 2018. Complete brood failure in an altricial bird is almost always associated with the sudden and permanent disappearance of a parent. *Journal of Animal Ecology* 87:1239–1250.
- Smith, C. C. and S. D. Fretwell, 1974. The optimal balance between size and number of offspring. *The American Naturalist* 108:499–506.
- Tinbergen, J. M. and S. Daan, 1990. Family planning in the great tit (*parus major*): optimal clutch size as integration of parent and offspring fitness. *Behaviour* 114:161–190.
- Tinbergen, J. M. and J. J. Sanz, 2004. Strong evidence for selection for larger brood size in a great tit population. *Behavioral Ecology* 15:525–533.
- Townsend, A. K., T. S. Sillett, N. K. Lany, S. A. Kaiser, N. L. Rodenhouse, M. S. Webster, and R. T. Holmes, 2013. Warm springs, early lay dates, and double brooding in a north american migratory songbird, the black-throated blue warbler. *PLoS One* 8:e59467.
- Tufto, J., 2015. Genetic evolution, plasticity, and bet-hedging as adaptive responses to temporally autocorrelated fluctuating selection: a quantitative genetic model. *Evolution* 69:2034–2049.
- Verhulst, S., J. Van Balen, and J. Tinbergen, 1995. Seasonal decline in reproductive success of the great tit: variation in time or quality? *Ecology* 76:2392–2403.
- Wesołowski, T., 2002. Anti-predator adaptations in nesting marsh tits *parus palustris*: the role of nest-site security. *Ibis* 144:593–601.
- Winkler, D. W., P. O. Dunn, and C. E. McCulloch, 2002. Predicting the effects of climate change on avian life-history traits. *Proceedings of the National Academy of Sciences* 99:13595–13599.
- Yamaguchi, N. and H. Higuchi, 2005. Extremely low nesting success and characteristics of life history traits in an insular population of *parus varius namiyei*. *The Wilson Journal of Ornithology* 117:189–194.

Supporting Information (SI) for

Multi-episodic fluctuating selection via fertility and viability in a great tit (*Parus major*) population

Implementation of the model with TMB

The implementation of our statistical model lies exclusively on the model-fitting technique named Template Model Builder (TMB, Kristensen et al., 2016). TMB is an R package for fast-fitting linear or nonlinear, Gaussian or non-Gaussian and any other statistical latent variable models. It provides great flexibility by allowing users to formulate the model in C++. It is strongly inspired by ADMB and thus is also characterized by Laplace approximation, which can yield marginal likelihood in the latent variable model. It requires some familiarity with C++ language but the templates are accumulating rapidly and can be accessed easily. The documentation of TMB is available at <http://kaskr.github.io/adcomp/index.html>.

It is worth noting that there is no built-in function for modeling VAR(1) processes in TMB. Therefore, we defined VAR(1) in C++ template ourselves in a way as introduced in Wei (2006) to avoid non-stationarity. Similarly, the zero-inflated beta-binomial function is also defined by us in C++ template. The parameter determining the zero-inflation probability can be turned off and disconnected from the rest of the model and the model thus is reduced to be a beta-binomial model, where the probability is a random variable with a beta distribution which gives rise to overdispersion.

Fitness components

In our statistical analysis of phenotypical selection, we take clutch size as the first fitness component. Some may argue that clutch size should not be taken as a fitness component, it's the mother's decision that how many eggs should be made. We claim that the number of fledglings and recruits for a given brood, which are usually taken as the measure of fecundity fitness is also dependent on the number of eggs the mother laid for the clutch, we thus include clutch size as the first episode to test if laying date has an effect on it. It is worth mentioning that excluding the first episode from the model would not affect the results of the other selective episodes.

We could have measured the individual female fitness as its local survival to next breeding season plus the half number of total recruits it reproduced in a given year as that in Sæther et al. (2016) from a population dynamics point of view. However, this would confound with the offspring fitness measure (y_3) and heavily complex the analysis in this study. We therefore simply take local mother survival as its fitness and understand this would potentially either overestimate or underestimate the fitness due to immigration and emigration. However, we believe that our local survival estimates are not biased by variation in recapture rates due to consistent high recapture rates for the studied population (Reed et al., 2013a).

Explanatory variables

Nicolaus et al. (2013) showed that females reduced clutch size with increasing population density. From a biological point of view, $\eta_{si}^{(\alpha)}$ might be affected by beech crop index (BCI, quantifies the number of beech nuts available in previous winter on a 3-point scale), breeding density (dens, annual number of first clutches) and clutch size (y_{0i}). The optimal laying date $\eta_{si}^{(\theta)}$ might be influenced by the spring temperature (temp) and the biomass peak date of caterpillars (peakd) (Chevin et al., 2015; Gienapp et al., 2013). However, spring temperature we used is highly correlated with caterpillar peak date in our study since the spring temperature data was obtained by climate window analysis screening out a climate window which correlates the food peak date best. Therefore, only one of them should be kept in our selected model. Clutch size may also influence the optimal laying date for a specific brood. The curvature of the stabilizing selection function $\eta_{si}^{(\omega)}$ is also likely being affected by the spring temperature. Data on the peak dates of caterpillar biomass are available from 1985 to 2013, while the great tit from 1955 to 2015. We find the annual caterpillar peak is well predicted by mean temperatures from 8th March to 17th May ($r^2 = 0.80$, Visser et al. (2006)). Therefore, we predict the data of caterpillar peak dates for the years where the data is absent with the spring temperature data.

The complete brood loss may be influenced by food resources available, for example, the late brood in a breeding season may starve to death. To test this, we include the relative laying date (RLD=laying date+30-food peak date, also defined as the individual mismatch for example in Reed et al. (2013a,b)) in the model. Besides, from a biological point of view, the breeding density determines the competition for food, therefore, it affects the caterpillar availability indirectly (see density-dependent compensation theory in Reed et al. (2013a)). It also might be influenced by predation which causes the sudden and permanent disappearance of one or both of the parents (Santema and Kempenaers, 2018). Since the timing of parents' disappearance was not recorded, instead, we estimate the difference in the annual sex-specific survival rate (the probability that a parent survives from year t to $t+1$, that is, was observed as a breeder in the following breeding season) between birds that hatched and did not hatch chicks, to investigate the effect of parental disappearance. In the analysis, we include only difference of females survival (abbreviated as diffsurvfem) since we find no difference in survival between the fathers hatched or not hatched broods. In addition, the finding from Cao et al. (2019) indicates that the mother effect is significant in explaining the variation of zero-inflation probability, so the mother effect as a random effect is also included in the model of zero-inflation probability. Julliard et al. (1997) found from a great tit population that in the year following peak nest predation, the mean clutch size was reduced by about one egg. We therefore also test if clutch size affects the probability of zero inflation since nest predation tends to be those of large clutches, as they are more noticeable to predators (frequent parental care, noisier offspring) (Slagsvold, 1982). Explanatory variables dens, temp, peakd, RLD, diffSurvFem were mean-centered since centering of input variables are simple means to improve the interpretability of regression coefficients (Schielzeth, 2010).

Model selection procedure

We fitted 87 candidate models in the model selection procedure, which can be found in R code in the supplementary material of this study. To we also fitted a list of models which is a shortcut of the 87 models to arrive at the best model efficiently. Table S1 only shows the shortcut of the candidate models from model A to model S. The candidate models are generally updated from simpler ones to the more complicated. Due to a large number of possibilities of candidate models, we followed a thumb rule to update candidate models that the significant variables in the previous model should be kept in the updated model. In Table S1, the column ΔAIC and Δp list the differences in AIC value and in the number of parameters kept in the models from the best model. The column Description describes how the subsequent model is updated from the previous one. Note that model A is a null model where only the eight parameters as shown in the table are turned on in the model.

Optimal clutch size

In the main text, we have found that no optimal clutch size exists in the sense of maximizing offspring viability from egg to recruit since the negative effect of clutch size on offspring survival from fledgling to recruit markedly dominates the effects in the other selective episodes. We claimed in the conclusion in the main text that this does not suggest that natural selection would favor small clutches since the number of recruits (or recruit value) is, in this case, a trade-off between clutch size and viability. To validate this claim, we show graphically the estimated and observed recruit value again clutch size as can be seen in Figure S1. We see from the figure that both the estimated (orange line) and observed recruit value (black line) generally increase along with bigger clutch size, excluding the grey shaded areas in which two few (less than 50, indicated by the sizes of points in the plot) brooding events fall in these clutch size groups. Therefore, the optimal clutch size suggested by the orange line is 13 and by observed recruit value, it is 12, both are much larger than the mean sample size in the whole analyzed data set, approximating to 9. The discrepancy between the optimal clutch size (either 12 or 13) and observed mean clutch size is potentially explained by the "bad year effect" in Boyce and Perrins (1987). Furthermore, either Pettifor et al. (1988) nor our analytical result suggest reproduction cost of breeding mothers in this bird species, therefore, the behind mechanism needs to be other investigated, which is out of the scope of this study.

Supplementary figure

Figure S2 shows plots of auto-correlation and cross-correlation function for the time series of observed population densities and spring temperatures. The plot of auto-correlation function for temperature shows significant non-zero auto-correlation with its 1, 2, 5, 7 lagged values but the correlations are weak. The top-right plot of auto-correlation for density indicates slight non-zero correlation at lag 1. Both of the plots in the bottom suggest weak but non-zero cross-correlation at 1 lag corr(dens_t, temp_{t-1}).

Model	ΔAIC	Δp	Description
A	3577.4	-51	$\alpha_{0,0} \neq \alpha_{0,1} \neq \alpha_{0,2} \neq \alpha_{0,3} \neq \alpha_{0,4} \neq 0;$ $\beta_{0,1'} \neq \beta_{0,2'} \neq \beta_{0,3'} \neq 0$
B	2783.02	-44	$\theta_{0,0} = \theta_{0,4} = 0, \theta_{0,1} \neq \theta_{0,2} \neq \theta_{0,3} \neq 0;$ $\omega_{0,0} = \omega_{0,4} = 0, \omega_{0,1} \neq \omega_{0,2} \neq \omega_{0,3} \neq 0;$ $\alpha_{layd}^{(1)} \neq 0$
C	2173.38	-31	$\alpha_{dens,0} \neq \alpha_{dens,1} \neq 0, \alpha_{dens,2} = \alpha_{dens,3} = \alpha_{dens,4} = 0;$ $\alpha_{BCI2,0} = \alpha_{BCI2,1} = 0, \alpha_{BCI2,2} \neq \alpha_{BCI2,3} \neq \alpha_{BCI2,4} \neq 0;$ Turn on BCI3 on $\eta_s^{(\alpha)}$ for all the episodes; $\omega_{temp,0} = \omega_{temp,4} = 0, \omega_{temp,1} \neq \omega_{temp,2} \neq \omega_{temp,3} \neq 0$
D	1960.84	-20	$\theta_{peakd,0} = \theta_{peakd,4} = 0, \theta_{peakd,1} \neq \theta_{peakd,2} \neq \theta_{peakd,3} \neq 0;$ $\theta_{dens,0} = \theta_{dens,4} = 0, \theta_{dens,1} \neq \theta_{dens,2} \neq \theta_{dens,3} \neq 0;$ $\beta_{clusize,1'} \neq \beta_{clusize,2'} \neq 0, \beta_{clusize,3'} = 0;$ $\beta_{dens,1'} \neq \beta_{dens,2'} \neq \beta_{dens,3'} \neq 0$
E	1936.92	-12	$\theta_{clusize,0} = \theta_{clusize,1} = \theta_{clusize,4} = 0, \theta_{clusize,2} \neq \theta_{clusize,3} \neq 0;$ $\beta_{peakd,1'} \neq \beta_{peakd,2'} \neq \beta_{peakd,3'} \neq 0;$ $\beta_{layd,1'} \neq \beta_{layd,2'} \neq \beta_{layd,3'} \neq 0$
F	1912.18	-12	$\beta_{diffSurvFem,1'} \neq \beta_{diffSurvFem,2'} \neq \beta_{diffSurvFem,3'} \neq 0;$ $\beta_{layd,1'} \neq \beta_{layd,3'} \neq 0, \beta_{layd,2'} = 0;$ $\beta_{dens,1'} = \beta_{dens,3'} = 0, \beta_{dens,2'} \neq 0$
G	1911.03	-13	$\alpha_{BCI3,0} \neq \alpha_{BCI3,1} \neq \alpha_{BCI3,3} \neq \alpha_{BCI3,4} \neq 0, \alpha_{BCI3,2} = 0;$
H	1655.94	-9	$\alpha_{year,0} = 0, \alpha_{year,1} \neq \alpha_{year,2} \neq \alpha_{year,3} \neq \alpha_{year,4} \neq 0;$
I	1039.32	-10	$\theta_{year,0} = \theta_{year,4} = 0, \theta_{year,1} = \theta_{year,2} = \theta_{year,3} \neq 0;$ $\theta_{peakd,1} = \theta_{peakd,2} \neq 0, \theta_{peakd,3} \neq 0;$ $\omega_{temp,0} = \omega_{temp,4} = 0, \omega_{temp,1} \neq 0, \omega_{temp,2} = \omega_{temp,3} \neq 0$
J	942.53	-11	$\beta_{year,1'} = \beta_{year,2'} \neq 0, \beta_{year,3'} \neq 0;$ $\beta_{diffSurvFem,1'} \neq 0, \beta_{diffSurvFem,2'} = \beta_{diffSurvFem,3'} = 0;$ $\beta_{layd,1'} \neq 0, \beta_{layd,2'} = \beta_{layd,3'} = 0$
K	790.55	-5	$\omega_{year,0} = \omega_{year,4} = 0, \omega_{year,1} \neq 0, \omega_{year,2} = \omega_{year,3} \neq 0;$ $\alpha_{year,0} = 0, \alpha_{year,1} \neq \alpha_{year,4} \neq 0, \alpha_{year,2} = \alpha_{year,3} \neq 0;$ $\phi_{\alpha,\alpha} = \phi_{\alpha,\theta} = \phi_{\alpha,\omega} = \phi_{\theta,\alpha} = 0, \phi_{\theta,\theta} \neq \phi_{\theta,\omega} \neq \phi_{\omega,\alpha} \neq \phi_{\omega,\theta} \neq \phi_{\omega,\omega} \neq 0$
L	498.31	3	$\alpha_{clusize,0} = \alpha_{clusize,4} = 0, \alpha_{clusize,1} \neq \alpha_{clusize,2} \neq \alpha_{clusize,3} \neq 0;$ $\alpha_{mother,0} \neq \alpha_{mother,1} \neq \alpha_{mother,2} \neq \alpha_{mother,3} \neq \alpha_{mother,4} \neq 0$
M	373.69	-1	$\beta_{peakd,1'} = \beta_{peakd,2'} \neq 0, \beta_{peakd,3'} \neq 0;$ $\omega_{temp,0} = \omega_{temp,4} = 0, \omega_{temp,1} = \omega_{temp,2} = \omega_{temp,3} \neq 0;$ $\beta_{mother,1'} \neq 0, \beta_{mother,2'} = \beta_{mother,3'} = 0;$ $\phi_{\alpha,\alpha} = \phi_{\alpha,\theta} = \phi_{\alpha,\omega} = \phi_{\theta,\alpha} = \phi_{\theta,\omega} = \phi_{\omega,\alpha} = \phi_{\omega,\theta} = 0, \phi_{\theta,\theta} \neq \phi_{\omega,\omega} \neq 0$
N	34.25	0	add random intercepts to the first episode $s = 0$ (clutch size is regressed on lay date)
O	24.59	-8	remove clutch size on $\eta_s^{(\theta)}$ for all the episodes; remove density on zip for all the episodes; $\alpha_{BCI2,0} = \alpha_{BCI2,1} = \alpha_{BCI2,2} = 0, \alpha_{BCI2,3} = \alpha_{BCI2,4} \neq 0;$ $\alpha_{BCI3,0} = \alpha_{BCI3,2} = 0, \alpha_{BCI3,1} \neq \alpha_{BCI3,3} \neq \alpha_{BCI3,4} \neq 0;$ $\omega_{0,0} = \omega_{0,4} = 0, \omega_{0,1} = \omega_{0,3} \neq 0, \omega_{0,2} \neq 0;$ $\beta_{peakd,1'} = \beta_{peakd,2'} \neq 0, \beta_{peakd,3'} = 0$
P	19.75	-9	$\beta_{0,1'} = \beta_{0,3'} \neq 0, \beta_{0,2'} \neq 0;$ $\theta_{peakd,0} = \theta_{peakd,4} = 0, \theta_{peakd,1} \neq \theta_{peakd,2} \neq \theta_{peakd,3} \neq 0;$ $\theta_{dens,0} = \theta_{dens,4} = 0, \theta_{dens,1} = \theta_{dens,2} \neq 0, \theta_{dens,3} \neq 0$
Q	18.65	-9	$\omega_{temp,0} = \omega_{temp,1} = \omega_{temp,4} = 0, \omega_{temp,2} = \omega_{temp,3} \neq 0$
R	16.55	-9	$\beta_{year,1} \neq 0, \beta_{year,2} = \beta_{year,3} \neq 0$
S	0	0	turn on all the entries in Φ ; $\beta_{peakd,1'} = \beta_{peakd,3'} = 0, \beta_{peakd,2'} \neq 0;$ $\alpha_{year,0} = 0, \alpha_{year,1} \neq \alpha_{year,2} \neq \alpha_{year,3} \neq \alpha_{year,4} \neq 0;$ $\alpha_{dens,0} \neq \alpha_{dens,2} \neq 0, \alpha_{dens,1} = \alpha_{dens,3} = \alpha_{dens,4} = 0;$ $\beta_{year,1'} \neq \beta_{year,2'} \neq \beta_{year,3'} \neq 0$

Table S1: A shortcut of model selection procedure. The candidate models were generally updated from simple one to complicated ones. The column ΔAIC and Δp list the difference in reported AIC value and number of parameters between the selected model and each candidate model respectively. The column Description gives the changes of each candidate model has based on the previous model (except for model A, where the Description simply describes the model itself). Subscripts again indicate the selection episode under consideration, i.e. 1 indicates $s = 1$ and 1' represents $s' = 1$, so as to the other episodes.

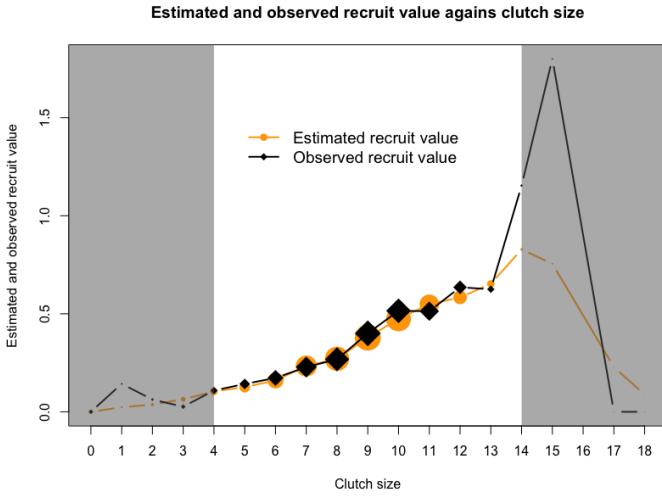


Figure S1: Estimated and observed recruit value against clutch size. The orange line with round-shaped points represents the estimated mean recruit value for each group of clutch size. The estimated recruit value for each brood in the analyzed data set is calculated by multiplying the true clutch size with non-zip and survival probability reported by our selected model $y_0(1 - \hat{zip}_1)(1 - \hat{zip}_2)(1 - \hat{zip}_3)\hat{p}_1\hat{p}_2\hat{p}_3$. The estimated recruit values then are grouped by clutch size and the mean recruit value for each clutch size group is calculated. The black line shows the observed mean recruit value for each clutch size group. The sizes of the points indicate the frequencies of observations in the analyzed data set. The grey shaded areas indicate the clutch size groups having less than 50 brooding records and the recruit values of which are not statistically powerful due to the small sample sizes.

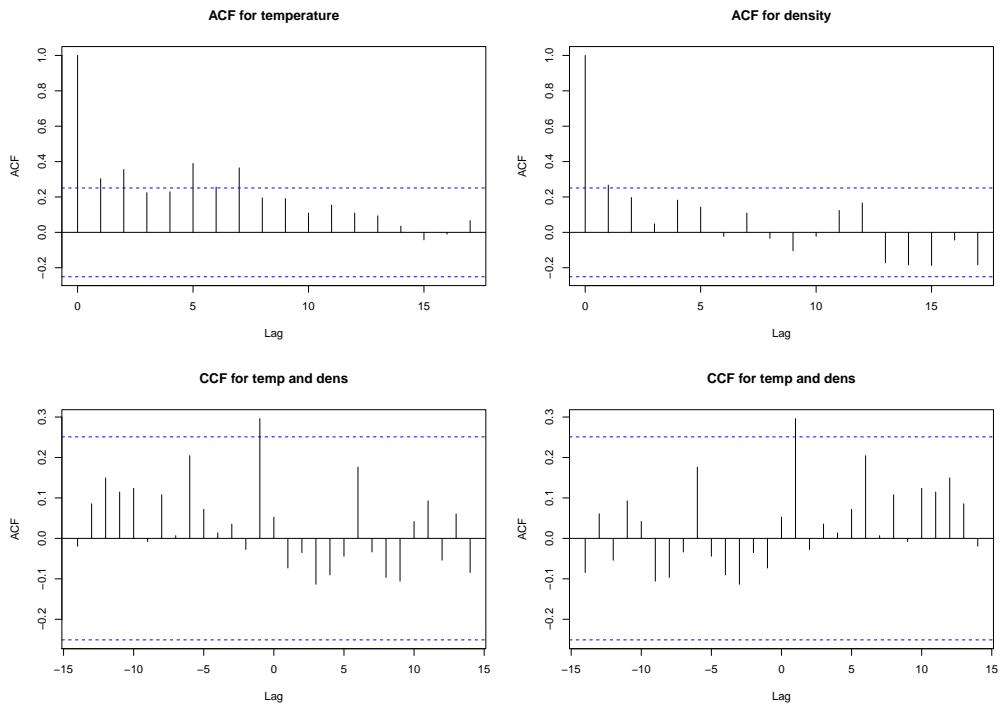


Figure S2: Auto-correlation and cross-correlation functions for the time series of observed population densities and temperatures. The bottom-left are the estimates of $\text{corr}(\text{temp}_t, \text{dens}_{t-k})$ and the bottom-right plot displays estimates of $\text{corr}(\text{dens}_t, \text{temp}_{t-k})$.

Literature Cited

- Boyce, M. S. and C. Perrins, 1987. Optimizing great tit clutch size in a fluctuating environment. *Ecology* 68:142–153.
- Cao, Y., M. E. Visser, and J. Tufto, 2019. A time-series model for estimating temporal variation in phenotypic selection on laying dates in a dutch great tit population. *Methods in Ecology and Evolution* 10:1401–1411.
- Chevin, L.-M., M. E. Visser, and J. Tufto, 2015. Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. *Evolution* 69:2319–2332.
- Gienapp, P., M. Lof, T. E. Reed, J. McNamara, S. Verhulst, and M. E. Visser, 2013. Predicting demographically sustainable rates of adaptation: can great tit breeding time keep pace with climate change? *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 368:20120289.
- Julliard, R., R. H. McCleery, J. Clobert, and C. M. Perrins, 1997. Phenotypic adjustment of clutch size due to nest predation in the great tit. *Ecology* 78:394–404.
- Kristensen, K., A. Nielsen, C. W. Berg, H. J. Skaug, and B. Bell, 2016. TMB: automatic differentiation and Laplace approximation. *Journal of Statistical Software* 70:1–21.
- Nicolaus, M., J. E. Brommer, R. Ubels, J. M. Tinbergen, and N. J. Dingemanse, 2013. Exploring patterns of variation in clutch size-density reaction norms in a wild passerine bird. *Journal of evolutionary biology* 26:2031–2043.
- Pettifor, R., C. Perrins, and R. McCleery, 1988. Individual optimization of clutch size in great tits. *Nature* 336:160.
- Reed, T. E., V. Grøtan, S. Jenouvrier, B.-E. Sæther, and M. E. Visser, 2013a. Population growth in a wild bird is buffered against phenological mismatch. *Science* 340:488–491.
- Reed, T. E., S. Jenouvrier, and M. E. Visser, 2013b. Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology* 82:131–144.
- Sæther, B.-E., M. E. Visser, V. Grøtan, and S. Engen, 2016. Evidence for r-and k-selection in a wild bird population: a reciprocal link between ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences* 283:20152411.
- Santema, P. and B. Kempenaers, 2018. Complete brood failure in an altricial bird is almost always associated with the sudden and permanent disappearance of a parent. *Journal of Animal Ecology* 87:1239–1250.
- Schielzeth, H., 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution* 1:103–113.
- Slagsvold, T., 1982. Clutch size variation in passerine birds: the nest predation hypothesis. *Oecologia* 54:159–169.
- Visser, M. E., L. J. Holleman, and P. Gienapp, 2006. Shifts in caterpillar biomass phenology due to climate change and its impact on the breeding biology of an insectivorous bird. *Oecologia* 147:164–172.
- Wei, W., 2006. *Time Series Analysis: Univariate and Multivariate Methods*. 2nd ed. Pearson Addison Wesley.

Bayesian inference with *tmbstan* for a state-space model with VAR(1) state equation

Yihan Cao,* Marcel E. Visser,† Jarle Tufto*

1 Introduction

Both frequentist and Bayesian statistical inference have been used for investigating ecological processes. In the frequentist framework, Template model builder (TMB, Kristensen et al., 2016), an R package developed for fast fitting complex linear or nonlinear mixed models, has gained the popularity recently, especially in the field of ecology which usually involves in modeling complicated ecological processes (for example Cadigan, 2015; Albertsen et al., 2016; Auger-Méthé et al., 2017). The combination of reverse-mode automatic differentiation and Laplace approximation for high-dimension integrals makes parameter estimation with TMB very efficient even for non-Gaussian and complex hierarchical models. TMB provides a flexible framework in model formulation and can be implemented even for statistical models where the predictor is nonlinear in parameters and random effect. However, the lack of capability of working in the Bayesian framework has hindered the adoption of it for Bayesians.

Within the Bayesian framework, the software package *Stan* (Gelman et al., 2015), a probabilistic programming language for statistical inference written in C++ attracts peoples attention. It uses the No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014), an adaptive extension to Hamiltonian Monte Carlo (Neal et al., 2011), which itself is a generalization of the familiar Metropolis algorithm, to conduct sampling more efficiently through the posterior distribution by performing multiple steps per iteration. Stan is a valuable tool for many ecologists utilizing Bayesian inference, particularly for problems where BUGS (Lunn et al., 2000) is prohibitively slow (Monnahan et al., 2017). As such, it can extend the boundaries of feasible models for applied problems, leading to a better understanding of ecological processes. Fields that would likely benefit include estimation of individual and population growth rates, meta-analyses and cross-system comparisons, among many others.

Combining the merits of TMB and Stan, the new software package *tmbstan* (Monnahan & Kristensen, 2018) which provides MCMC sampling for TMB models was developed. This package provides ADMB and TMB users a possibility for making Bayesian statistical analysis when prior information on the unknown parameters is available. From the user's perspective, it implements NUTS sampling from a target density proportional to the product of marginal likelihood (computed by TMB or Stan) and the prior density

*Centre for Biodiversity Dynamics, Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway

†Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Post Office Box 50, 6700AB Wageningen, Netherlands

specified by the user. The user has the flexibility to decide which random effects are integrated out via the Laplace approximation in TMB and then the TMB model is passed to function Stan in the RStan package so that the rest of the parameters are integrated by Stan. This methodology might therefore potentially be more computationally efficient than using MCMC alone to integrate out all parameters. Monnahan and Kristensen (2018) introduced the *tmbstan* package, applied it to simulation studies and compared its capabilities (computational efficiency and the accuracy of Laplace approximation) with other platforms such as ADMB and TMB.

However, it is unclear that if Bayesian inference with arbitrary prior distribution implemented with Stan would perform comparatively with frequentist inference when modeling complex ecological processes. It is also unclear that when using *tmbstan*, if using the Laplace approximation to integrate latent variables is more computationally efficient than handling all latent variables via MCMC. In the case studies in Monnahan and Kristensen (2018), Laplace approximation turned out to reduce the computational efficiency of MCMC. Another issue arose in the case studies is that the Laplace approximation to the integration of random effects is not accurate to a degree and this could lead to biased parameter estimates or uncertainties in parameter estimation. To gain more insights on these issues, in this paper we conduct simulation studies and a case study in the context of modeling fluctuating and auto-correlated selection with state-space models (SSM). These forms of models are more generally increasingly used in ecology to model time-series such as animal movement paths and population dynamics (for example Cadigan, 2015; Albertsen et al., 2016; Auger-Méthé et al., 2017). Furthermore, following Cao, Visser, and Tufto (2019), we also use order-1 vector autoregressive model (VAR(1)) to model the unobserved states, which in our study are temporally fluctuating and potentially auto-correlated height, width and location of a Gaussian fitness function. This also allows us to make a further investigation into the issue of underestimation of the auto-correlation parameter in auto-regressive models shown in Chevin et al. (2015) and Cao et al. (2019).

Through the simulation and empirical studies, our paper aims to (1) compare estimates between frequentist inference and Bayesian inference under different simulation schemes; (2) investigate how the choice of prior influence Bayesian inference; (3) compare the computational efficiency of MCMC with and without integrating out some of the random effects via Laplace approximation.

2 Methodology

2.1 Model formulation

We consider a typical ecological process, the fluctuating selection in a bird species, the great tit (*Parus major*). We conduct the study in the context of temporally changing selection on the laying date with the number of fledglings as the fitness component, but it can be generalized to any episode of viability or fertility selection, or to overall selection through lifetime fitness. The discrete nonnegative variable, number of fledglings, is best modelled by distributions such as Poisson, or zero-inflated Poisson (for example Chevin et al., 2015; Cao et al., 2019). Within the framework of generalized linear models, the expected value of response variable is commonly linked to the linear predictors of biologically interest by logarithm. When both linear and quadratic effects of the traits are included, this leads to a Gaussian model of stabilizing selection. In this study, the number

of fledglings in a specific brood is assumed to be Poisson distributed, $X_i|w_i \sim \text{Poisson}(w_i)$, where i indicates the breeding event. The fitness (the expected number of fledglings w_i) of individuals with phenotype z_i is then given by

$$\ln w_i = \eta_t^{(\alpha)} - \frac{(z_i - \eta_t^{(\theta)})^2}{2(e^{\eta_t^{(\omega)}})^2}, \quad (1)$$

where $\eta_t^{(\alpha)}$, $\eta_t^{(\theta)}$ and $e^{\eta_t^{(\omega)}}$ (e based to guarantee positive) are parameters determining the logarithm of maximum fitness, optimum laying date and width of the fitness function in year t respectively. We further model $\eta_t^{(\alpha)}$, $\eta_t^{(\theta)}$ and $\eta_t^{(\omega)}$, the three stochastic processes as following:

$$\begin{aligned} \eta_t^{(\alpha)} &= \mu_\alpha + \sigma_\alpha \alpha_t, \\ \eta_t^{(\theta)} &= \mu_\theta + \sigma_\theta \theta_t, \\ \eta_t^{(\omega)} &= \mu_\omega + \sigma_\omega \omega_t. \end{aligned} \quad (2)$$

The elements of vector $\mu = (\mu_\alpha, \mu_\theta, \mu_\omega)^T$ are the means of the three processes. The stochastic processes $\alpha_t, \theta_t, \omega_t$ are assumed to be multivariate normal distributed $(\alpha_t, \theta_t, \omega_t)^T \sim \mathbf{N}_3(\mathbf{0}, \boldsymbol{\Gamma}_0)$ with $\boldsymbol{\Gamma}_0 = \begin{bmatrix} 1 & \rho_{\alpha,\theta} & \rho_{\alpha,\omega} \\ \rho_{\alpha,\theta} & 1 & \rho_{\theta,\omega} \\ \rho_{\alpha,\omega} & \rho_{\theta,\omega} & 1 \end{bmatrix}$, where $\rho_{\alpha,\theta}$, $\rho_{\alpha,\omega}$ and $\rho_{\theta,\omega}$ indicate the correlations and are assumed to be mutually independent. $(\alpha_t, \theta_t, \omega_t)^T$ are further assumed to follow a first-order vector autoregressive (VAR(1)) process as below:

$$\begin{bmatrix} \alpha_t \\ \theta_t \\ \omega_t \end{bmatrix} = \boldsymbol{\Phi} \begin{bmatrix} \alpha_{t-1} \\ \theta_{t-1} \\ \omega_{t-1} \end{bmatrix} + \mathbf{w}_t, \quad (3)$$

where $\boldsymbol{\Phi}$ is 3×3 transition matrix and \mathbf{w}_t is a 3-dimentional vector of white noise. The covariance matrix of \mathbf{w}_t is calculated as $\boldsymbol{\Gamma}_0 - \boldsymbol{\Phi}\boldsymbol{\Gamma}_0\boldsymbol{\Phi}$. Correlations between the elements of \mathbf{w}_t are determined by both $\rho = (\rho_{\alpha,\theta}, \rho_{\alpha,\omega}, \rho_{\theta,\omega})$ and $\boldsymbol{\Phi}$. If ρ is $\mathbf{0}$ vector and $\boldsymbol{\Phi}$ is diagonal, then \mathbf{w}_t reduces to be three independent and identically distributed white noise processes. In this case, α_t , θ_t and ω_t simplify to three independent first-order autoregressive (AR(1)) processes. If ρ is $\mathbf{0}$ and all entries of $\boldsymbol{\Phi}$ are zero, both $(\alpha_t, \theta_t, \omega_t)^T$ and \mathbf{w}_t reduce to three independent and identically distributed white noise processes. In any case, our non-centered parameterization implies that the standard deviation of α_t , θ_t and ω_t is only determined by σ_α , σ_θ and σ_ω respectively. We expect the non-centered parameterization yields simpler posterior geometries (Betancourt & Girolami, 2015) and will be much more efficient in terms of effective sample size when there is not much data (Stan Development Team, 2018b, chapter 20).

It is worth mentioning that one objective of this study is to provide another case study beyond the ones in Monnahan and Kristensen (2018). Therefore, even though α_t , θ_t and ω_t are assumed to be VAR(1) in the model, in the simulation study we consider only AR(1) θ_t and white noise of α_t and ω_t . The alternative simulation studies in which α_t , θ_t and ω_t are formulated as other possible stochastic processes can be conducted similarly and exhaustively, but that is an enormous amount of work in one single study. When α_t , θ_t and ω_t are assumed to be VAR(1), one caution to be taken is that all the eigenvalues of $\boldsymbol{\Phi}$ must lie in the unit circle to guarantee the VAR (1) process to be stationary (Wei, 2006). At last, in the simulation study, we assume that the model structure is known, which means that we already know θ_t is AR(1) process since the aim of the study is not to explore the structure of the true model.

2.2 Prior distribution

The priors are assumed to be independent to each other $\pi(\mu, \Phi, \Sigma) = \pi(\mu)\pi(\Phi)\pi(\Sigma)$. We take a normal $N(\mathbf{m}, q\mathbf{I}_3)$ prior distribution for the process mean vector $\mu = (\mu_\alpha, \mu_\theta, \mu_\omega)$ and input weak prior information on the process mean by taking $\mathbf{m} = \mathbf{0}$ and $q = 100$. Since in this study we assume constant $\eta_t^{(\alpha)}$ and $\eta_t^{(\omega)}$, $\phi_{\theta,\theta}$ is the only non-zero entry in Φ . We used truncated normal prior on $\phi_{\theta,\theta}$ since it outperforms Jeffreys' prior (Jeffreys & Jeffreys, 1961), g prior (Zellner, 1986) and natural conjugate prior (Schlaifer & Raiffa, 1961) in terms of posterior sensitivity using Highest Posterior Density Region (HPDR) criterion concluded from the simulation study in Karakani et al. (2016). Lei et al. (2011) also uses truncated normal distribution as subjective prior for the auto-regressive parameter in its AR (1) model. The mean and standard deviation of the truncated normal distribution are arbitrarily set to be 0 and 0.5 respectively.

For the variance of the error term σ_θ^2 (σ_α^2 and σ_ω^2 are assumed to be zero), two priors are used:

- (1) half-Cauchy (0, 10) prior on σ_θ (Prior1);
- (2) lognormal (1, 0.5) prior on σ_θ (Prior2).

These two priors are referred to Prior1 and Prior2 respectively in the rest of this paper. It is worth mentioning that we also tested uniform prior on $\log(\sigma_\theta)$ (non-informative improper prior which equals to $1/\sigma$ prior on σ (Gelman et al., 2006)) and inverse-gamma (1, 1) prior on σ_θ^2 (non-informative proper prior, also illustrated in (Gelman et al., 2006)), but both of them render an issue that the sampler traps in a subspace of the whole parameter space of $\log(\sigma_\theta)$ and results in numerous divergent transitions. It was potentially caused by the posterior becoming improper and consisting of a mode and an infinite low-posterior-density ridge extending to infinity as illustrated in Tufto et al. (2012). We thus in this study only consider the two proper informative priors (Prior1 and Prior2), while more information on the MCMC with inverse-gamma (1, 1) prior on σ_θ^2 is given in Supporting Information.

Note also that the scale parameters $\log(\sigma_\theta)$ is declared in the TMB template in the logarithmic format, but the half-Cauchy prior and lognormal prior contributed to the total likelihood with the log density in terms of σ_θ and for inverse-gamma prior, it is in terms of σ_θ^2 , where σ_θ is a positive transform $\sigma = e^{\log\sigma}$. Therefore, Jacobian adjustment (see chapter 20.3 in Stan Development Team (2018b) for Jacobian adjustment) was conducted by adding $\log\sigma_\theta$ to the total likelihood when half-Cauchy prior and lognormal prior are used. When testing inverse-gamma prior, it was $\log 2 + 2 \log \sigma_\theta$ added to the total likelihood.

2.3 Software implementation

The model is formulated with C++ and passed to TMB for frequentist inference. The model objective (fn) and gradient (gr) functions are fed to optimization function *nlnminb* with default setting to optimize the objective function.

For Bayesian inference, the TMB model objective and gradient functions are passed to *tmbstan* which uses the *stan* function and executes the No-U-Turn sampler (NUTS) algorithm by default to sample. Currently the other options are "HMC" (Hamiltonian Monte Carlo), and "Fixed_param". We ran the simulation study on a multicore computing server with enough RAM to avoid swapping to disk. The number of warmup iterations to be excluded when computing the summaries is set to 1000 and for total sample length, it is 3000. We thin each chain to every second sample and set the value

adapt_delta to 0.95, which is the average proposal acceptance probability Stan aims for during the adaption (warmup) period. We set a seed for each simulation including data set and tmbstan to make sure all the simulation results are reproducible.

Divergent transitions during sampling may occur due to a large step size in the sampler or a poorly parameterized model, meaning that the iteration of the MCMC sampler runs into numerical instabilities (Carpenter et al., 2017) and thus inferences will be biased. RStan team suggested that the problem may be alleviated by increasing the adapt_delta parameter (gives a smaller step size), especially when the number of divergent transitions is small (Stan Development Team, 2018a). In our simulation studies, we find it difficult to completely avoid divergent transitions across all data sets even though adapt_delta is increased to 0.95. Similar to Fuglstad, Hem, Knight, Rue, and Riebler (2019), we thus removed simulations where 0.1% or more divergent transitions in the iterations after warmup occur during the inference to avoid reporting biased results.

It is worth mentioning that the execution of Markov chains can be done in parallel. While the default of RStan is to use 1 core, the RStan team recommended to set it to as many processors as the hardware and RAM allow and at most one core per chain (Stan Development Team, 2018a). The simulations we run are done with a server that has 28 available cores. We thus set the number of cores to be 4 for the 4 Markov chains. However, since for frequentist inference, optimization algorithm used in R function "nlsminb" makes the best use of all available cores of CPU, we thus only compare the computational efficiency between tmbstan with and without Laplace approximation and ignore the computational efficiency with "nlsminb" to ensure fair comparisons.

3 Simulation scheme and results

3.1 Simulation scheme

All the data simulated are in natural units and considered to be biologically realistic according to the empirical studies of natural birds populations (e.g. Grant & Grant, 2002; Vedder et al., 2013). Samples were modeled from a population undergoing stabilizing selection with AR(1) θ_t , fixed $\eta_t^{(\alpha)}$ and $\eta_t^{(\omega)}$. Vector $\mu = (\mu_\alpha, \mu_\theta, \mu_\omega)^T$ is set to (2, 20, 3.5). The autocorrelation $\phi_{\theta,\theta}$ is set to 0.1, 0.4 and 0.7 (only positive values considered since the estimate of auto-correlation in temporal optimal laying date is positive, for example 0.3029 in Chevin et al. (2015) and 0.524 in Cao et al. (2019)), the variance of fluctuating optimal laying date σ_θ is set to 20.

For each value of $\phi_{\theta,\theta}$, $tmax = 25$ or 50 time points were simulated and for each time point the sample size was drawn from a Poisson distribution with mean $n = 25, 50$ or 100 individuals. We considered four combinations of $tmax$ and n , which are ($tmax = 25, n = 50$), ($tmax = 25, n = 100$), ($tmax = 50, n = 25$) and ($tmax = 50, n = 100$). These four combinations are referred as simulation setting 1, 2, 3, 4 respectively in the following sections. Similar to Cao et al. (2019), we neglected response to selection and used the same normal distribution for simulating individual phenotype each year. The phenotypic standard deviation before selection σ_z was set to 20, such that the strength of stabilizing selection $S = \sigma_z^2/(e^{\eta_t^{(\omega)}})^2 + \sigma_z^2$ (e.g. Chevin et al., 2015) was 0.267. For each individual, its fitness was computed from its phenotype using equation (1), and its actual number of offspring was then drawn from a Poisson distribution with mean $w_t(z)$.

3.2 Frequentist vs. Bayesian estimates

The results of one single simulation obtained from maximum likelihood in the frequentist framework are compared with those from *tmbstan*. The summaries of the estimates with *tmbstan* are computed after dropping the warmup iterations and merging the draws from all the four chains. The frequentist and Bayesian estimates with different sample sizes and $\phi_{\theta,\theta} = 0.4$ are shown in Table 1, the estimates with other values of auto-correlation in θ_t ($\phi_{\theta,\theta} = 0.1$ and 0.7) can be found in Supporting Information.

From Table 1 we find that both frequentist and Bayesian inferences show good estimates for μ_α and μ_ω . It is interesting to see that the auto-correlation for θ_t is not always under-estimated under all settings (for example ($tmax = 25, n = 50$)), this can be also seen from the tables for parameter estimates in Supporting Information. Bayesian inference with Prior1 (half-Cauchy prior) generally reports smaller estimates of μ_θ than MLE and Prior2 (lognormal prior) but larger estimates of $\phi_{\theta,\theta}$ and $\log\sigma_\theta$. The estimates with MLE and Prior2 are close to each other while the estimates with Prior2 show fewer uncertainties for $\phi_{\theta,\theta}$ and $\log\sigma_\theta$ implied by the smaller standard errors in the brackets. Prior2 also reports smaller estimates for $\log\sigma_\theta$ compared with MLE and Prior1 since it puts very large weight on small values of the variance, as will be graphically demonstrated in section 3.4. We also find that $\phi_{\theta,\theta}$ and $\log\sigma_\theta$ are difficult parameters to estimate since none of these three techniques can estimate them accurately across all the cases. However, the estimates are based on one realization of simulation, the discrepancy between estimates to the true value would vary from simulation to simulation.

We also compare the estimates across the different sample sizes. We typically compare the estimates between setting ($tmax = 25, n = 50$) and ($tmax = 25, n = 100$), ($tmax = 50, n = 25$) and ($tmax = 50, n = 100$), ($tmax = 25, n = 100$) and ($tmax = 50, n = 100$). We find that increasing the mean sample size at each time point does not necessarily increase the certainty of the estimates, but the data set with increased time points ($tmax = 50, n = 100$) contains more information on the parameters of interest and thus reports more certain estimates compared with the data set with ($tmax = 25, n = 100$). The same conclusion can be also drawn by making similar comparisons among the estimates in Table S1 and S2 in Supporting Information.

We can also find from Table 1, Table S1 and S2 from Supporting Information that the Bayesian inference with Prior1 in some cases report 1 or 2 divergent transitions while with Prior2 there are no divergent transitions reported. This implies that the geometric shape of posterior likelihood with Prior1 is more challenging for sampling probably due to light tails and thus potentially leads to an incomplete exploration of the target distribution.

3.3 Bias Plot

The comparison between the estimates in the last section is based on one realization of the simulation. To make comparisons of estimates over more realizations, the simulation was repeated 50 times under the setting of ($tmax = 50, n = 25$). Due to divergent transitions, only 44 out of 50 replicates were kept and the replications with more than 0.1% divergent transitions (in 2000 iterations) were excluded from the analysis. For the estimate of $\phi_{\theta,\theta}$ and $\log\sigma_\theta$ in each replication, the bias was calculated in a frequentist framework as the absolute difference between the true value and the mean estimate from each inference technique. The absolute bias for $\phi_{\theta,\theta}$ and $\log\sigma_\theta$ are graphically displayed in the upper and lower plot in Fig. 1 respectively. From the upper plot we find that in most replications, Bayesian inference with Prior1 slightly outperforms the frequentist

Table 1: Frequentist and Bayesian estimates (standard errors) from the model with AR(1) θ_t , autocorrelation in θ_t $\phi_{\theta,\theta} = 0.4$, and different sample sizes (($tmax = 25, n = 50$), ($tmax = 25, n = 100$), ($tmax = 50, n = 25$) and ($tmax = 50, n = 100$)) from one realization of the simulation. For each sample size setting, the number of divergent transitions in the MCMC is also reported and is used as a measure of stability of the inference scheme. MLE stands for maximum likelihood estimate, Prior1 and Prior2 represent half-Cauchy (0, 10) and lognormal (1, 0.5) prior respectively.

$\phi_{\theta,\theta} = 0.4, tmax = 25, n = 50$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	1	0
μ_α	2	2.017(0.015)	2.017(0.015)	2.016(0.015)
μ_θ	20	18.5(3.7)	18.3(5.1)	18.5(3.7)
μ_ω	3.5	3.472(0.028)	3.475(0.028)	3.469(0.028)
$\phi_{\theta,\theta}$	0.4	0.14(0.20)	0.23(0.23)	0.16(0.18)
$\log\sigma_\theta$	2.996	2.77(0.15)	2.88(0.19)	2.70(0.14)
$\phi_{\theta,\theta} = 0.4, tmax = 25, n = 100$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	2	0
μ_α	2	1.995(0.011)	1.995(0.012)	1.995(0.012)
μ_θ	20	20.2(8.7)	18.3(17.5)	20.1(7.4)
μ_ω	3.5	3.506(0.022)	3.508(0.022)	3.504(0.021)
$\phi_{\theta,\theta}$	0.4	0.50(0.17)	0.59(0.18)	0.46(0.13)
$\log\sigma_\theta$	2.996	3.25(0.18)	3.43(0.28)	3.13(0.14)
$\phi_{\theta,\theta} = 0.4, tmax = 50, n = 25$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	0	0
μ_α	2	1.974(0.015)	1.974(0.015)	1.973(0.015)
μ_θ	20	20.0(3.8)	19.8(4.9)	20.1(4.2)
μ_ω	3.5	3.520(0.032)	3.523(0.032)	3.515(0.031)
$\phi_{\theta,\theta}$	0.4	0.42(0.14)	0.48(0.15)	0.42(0.13)
$\log\sigma_\theta$	2.996	2.84(0.13)	2.92(0.16)	2.79(0.13)
$\phi_{\theta,\theta} = 0.4, tmax = 50, n = 100$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	0	0
μ_α	2	1.9865(0.0076)	1.9864(0.0076)	1.9861(0.0076)
μ_θ	20	20.7(3.9)	20.0(5.0)	20.7(4.1)
μ_ω	3.5	3.512(0.015)	3.513(0.015)	3.510(0.015)
$\phi_{\theta,\theta}$	0.4	0.41(0.13)	0.47(0.15)	0.41(0.12)
$\log\sigma_\theta$	2.996	2.89(0.12)	2.97(0.17)	2.85(0.11)

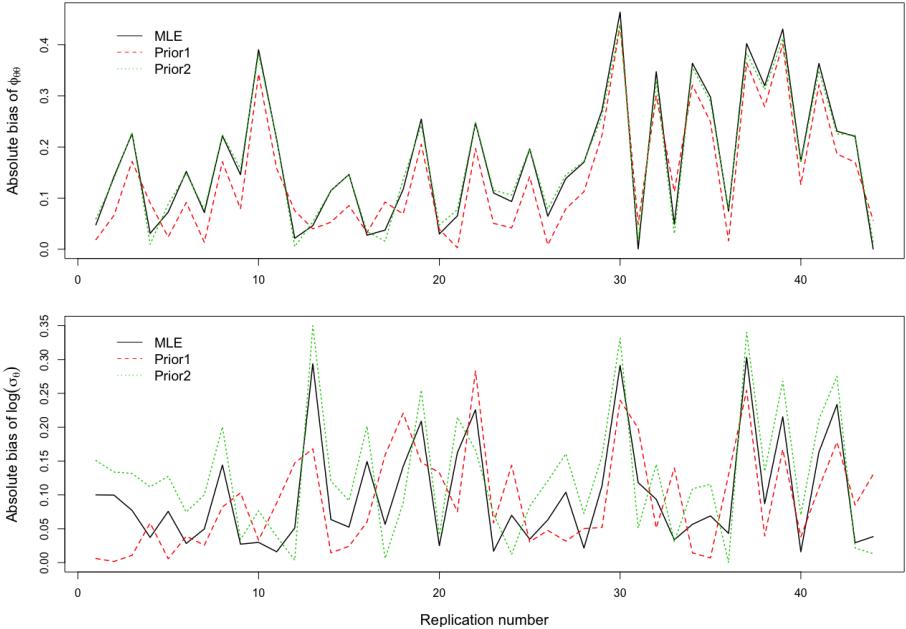


Figure 1: Bias plots for the auto-regressive parameter $\phi_{\theta,\theta}$ (the upper plot) and for the scale parameter $\log\sigma_\theta$ (the lower plot) respectively under the setting with time series length $tmax = 50$, average annual sample size $n = 25$, autocorrelation in $\theta_t \phi_{\theta,\theta} = 0.4$ and 44 replications (50 replications were conducted, among which 6 replications report 3 or more divergent transitions for the MCMC of Bayesian inference and thus are removed from the analysis).

inference and Bayesian inference with Prior2, the latter two reported very close estimates for $\phi_{\theta,\theta}$. One striking thing is that the bias is close to or even larger than 0.4 for some replications, this suggests that the inferences report even negative estimates of $\phi_{\theta,\theta}$ and it again turns out to be a difficult parameter. In the lower plot, we can see no single inference technique stands out in estimating the scale parameter $\log\sigma_\theta$.

3.4 Prior-posterior distribution

Fig. 2 shows histograms of posterior samples of the scale parameter σ_θ from models with the two different prior distributions: half-Cauchy (0, 10) and log-normal (1, 0.5), which are represented by solid lines in the left and right plot on each subplot respectively. The true value of σ_θ is indicated by a solid red line. Plot (a), (b), (c) and (d) correspond to setting $(tmax = 25, n = 50)$, $(tmax = 25, n = 100)$, $(tmax = 50, n = 25)$ and $(tmax = 50, n = 100)$ respectively. We can see from plot (a) that the priors are quite informative and pull the posteriors towards small values away from the true value and this prior-domination is more clear with log-normal prior where the prior distribution sharply peaks at 2. The domination is not mitigated even though the mean annual sample size is increased to 100 as shown in plot (b). With the same total sample size in plot (c) ($tmax = 50, n = 25$) as that in plot (a) ($tmax = 25, n = 50$), the posterior likelihoods in plot (c) are, however, not dominated by the priors. The prior-domination is also mitigated in plot (d) compared with plot (b) by increasing the time points from

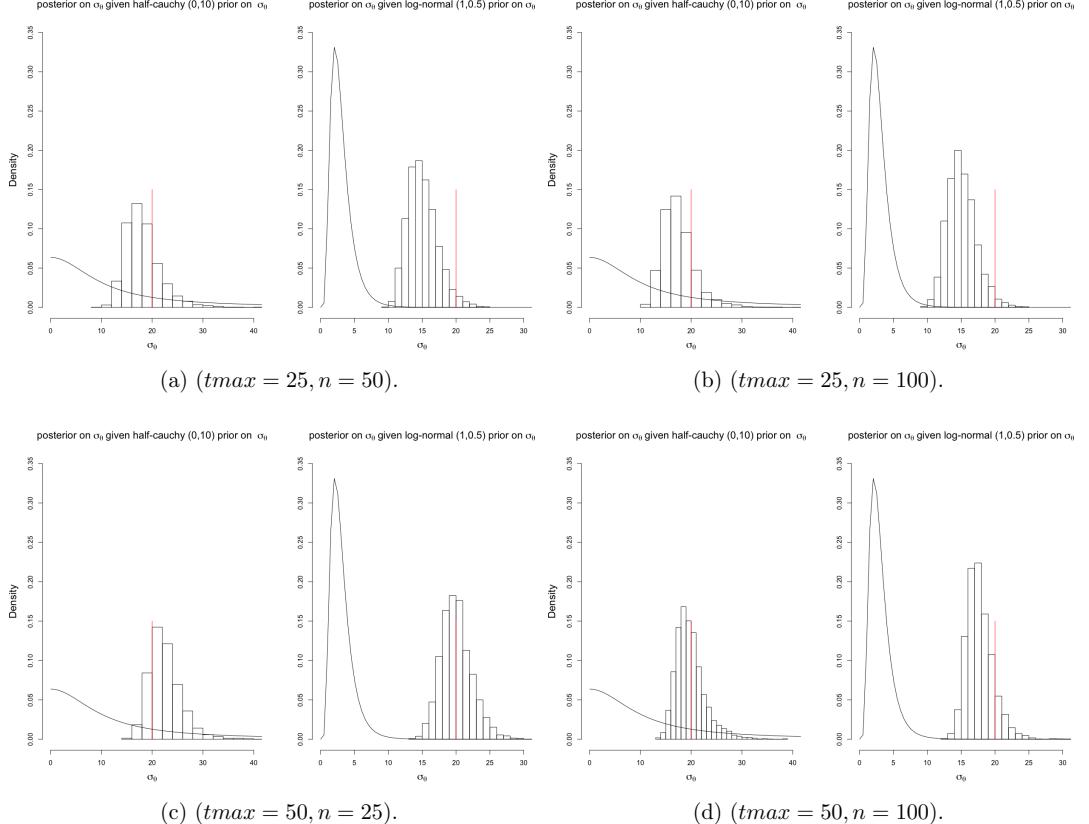


Figure 2: Histograms of posterior samples of the scale parameter σ_θ from models with two different prior distributions. Plot (a), (b), (c) and (d) correspond to sample size setting ($t_{max} = 25, n = 50$), ($t_{max} = 25, n = 100$), ($t_{max} = 50, n = 25$) and ($t_{max} = 50, n = 100$) respectively. On each subplot, the left one shows the histogram of posterior samples given half-Cauchy (0, 10) prior on σ_θ and similarly, the right one displays the histogram of posterior samples given log-normal (1, 0.5) prior on σ_θ . Overlaid on each subplot (the solid black lines) is the corresponding prior density function. The red lines indicate the true value of σ_θ . Only $\phi_{\theta,\theta} = 0.4$ was considered in the simulations.

25 to 50.

Altogether, the informative log-normal prior pulls more of the posterior towards a narrower range of smaller parameter values especially when the number of time points in the data is small. The posterior samples are less dominated by the half-Cauchy prior in this case. Increasing the annual mean sample size does not necessarily lead to better identification of the small region of parameter space. Only the amount of time points is the matter for the likelihood to overwhelm the prior distribution and to dominate the posterior distribution.

3.5 Computational efficiency with and without Laplace approximation

In *tmbstan*, sampling can be performed with or without Laplace approximation for the random effects. It is possible to mix the Laplace approximation with MCMC by specifying *laplace=TRUE*, such that the random effects are integrated with the Laplace approximation in TMB and other parameters (such as fixed effects and hyperparameters specifying the distribution of the random effects) are handled by the NUTS in Stan. In the case studies in Monnahan and Kristensen (2018), the Bayesian inference algorithms with Laplace approximation is less computationally efficient than without Laplace approximation, where the efficiency is defined as the minimum effective sample size per second. Following that definition, we calculated the efficiency of *tmbstan* with and without Laplace approximation with simulated data. Different from Monnahan and Kristensen (2018), we did not consider the computational efficiency of Frequentist inference with the Laplace approximation, as explained in the last section.

In Fig. 3, plot (a) displays violin plots of computational efficiency without (orange) and with (green) Laplace approximation (la) of Bayesian inference with Prior1 under different sample size settings. The setting 1, 2, 3, 4 on x axis stand for setting ($tmax = 25, n = 50$), ($tmax = 25, n = 100$), ($tmax = 50, n = 25$) and ($tmax = 50, n = 100$) respectively. Only $\phi_{\theta,\theta} = 0.4$ was considered and the divergent transitions were not taken into account. Inside the violin plots are box plots showing the quantiles of 50 realized computational efficiencies. Similarly, the violin plots of computational efficiency with Prior2 are shown on plot (b). We find from both plot (a) and (b) that Bayesian inference without Laplace approximation generally is more efficient under setting 1, 2, and 3, the outperformance is more manifest when the sample size is small ($tmax = 25, n = 50$). However, when the sample size is increased to ($tmax = 50, n = 100$), inference with Laplace approximation turns out to be slightly more efficient than that without Laplace approximation, the boxplots and violin plots also tend to be more compact under this setting.

Even though the technique in which the random effects are integrated out by Laplace approximation in TMB turns out to be less efficient in most settings, we still provide a counterexample from Monnahan and Kristensen (2018) in which the enabling of Laplace approximation is always less computationally efficient in the case studies.

3.6 Laplace approximation check

By comparing the Bayesian posteriors with and without Laplace approximation, we are allowed to check how well the Laplace approximation works. Fig. 4 shows pair plots of posterior samples with and without Laplace approximation done by TMB under different sample size settings with Prior2. Only autocorrelation in $\theta_t \phi_{\theta,\theta} = 0.4$ was considered.

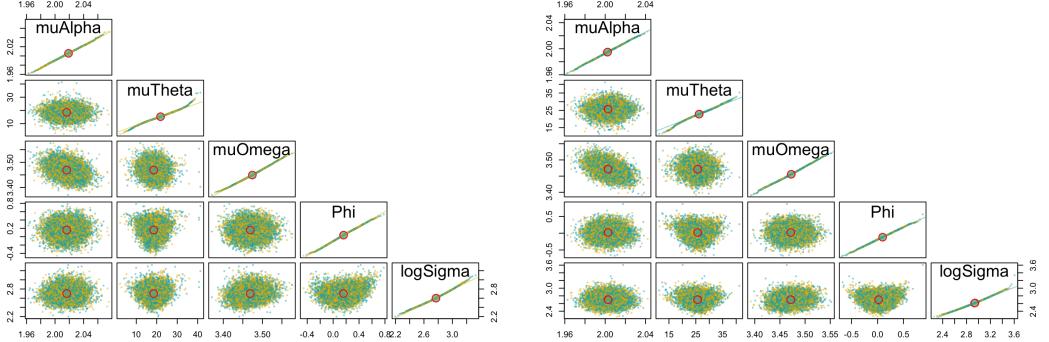


(a) Computational efficiency with Prior1.

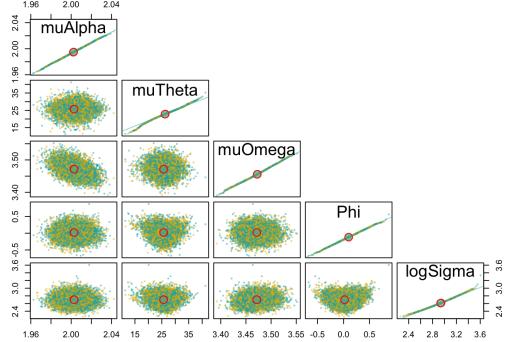


(b) Computational efficiency with Prior2.

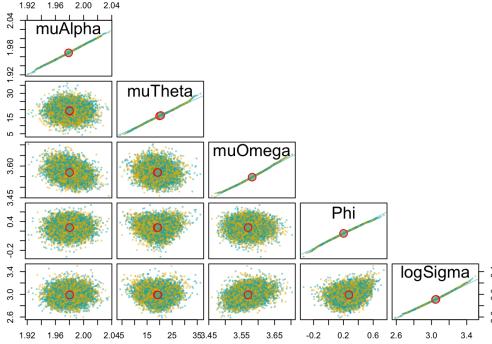
Figure 3: Violin plots of computational efficiency (minimum effective sample size per second) without (orange) and with (green) Laplace approximation (la). The four settings on x axis correspond to sample size setting ($tmax = 25, n = 50$), ($tmax = 25, n = 100$), ($tmax = 50, n = 25$) and ($tmax = 50, n = 100$) respectively. Plot (a) shows the computational efficiency of Bayesian inference with Prior1 and plot (b) with Prior2. Only $\phi_{\theta,\theta} = 0.4$ was used in simulations. Inside the violin plots are box plots showing the quantiles of 50 realized computational efficiencies. For each realization among the 50 simulations and across the settings, the same specifications in `tmbstan` are used.



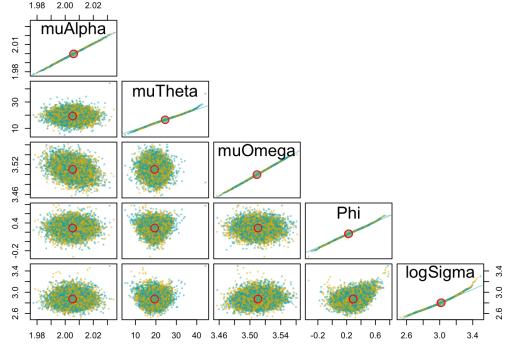
(a) ($t_{max} = 25, n = 50$) with Prior2.



(b) ($t_{max} = 25, n = 100$) with Prior2.



(c) ($t_{max} = 50, n = 25$) with Prior2.



(d) ($t_{max} = 50, n = 100$) with Prior2.

Figure 4: Pair plots of posterior samples for Laplace approximation check from one realization of the simulation with Prior2. The four plots (a) (b) (c) and (d) correspond to the four settings of sample size in simulation. The random effects in the TMB model can be integrated with two techniques: (1) full MCMC integration via NUTS and (2) Laplace approximation. To check the accuracy of Laplace approximation to the posterior likelihood density, the posterior samples for all the fixed effects in the model without (yellow dots) and with Laplace approximation (green dots) are shown pair-wisely on the same plot. Columns and rows on the lower diagonal correspond to pair-wise parameters, with the diagonal showing QQ-plot of posterior samples from Bayesian inference without (yellow dots) and with (green dots) Laplace approximation for that parameter including a 1:1 line in yellow. The large red circles on the off-diagonal plots represent the pairwise means. On each off-diagonal plot, there are 4000 yellow dots corresponding to 1000 samples retained from each of four chains without Laplace approximation, so as the green dots with Laplace approximation. Posterior rows were randomized to prevent consistent overplotting of one integration technique. Overlaps in the two colored dots suggest that the Laplace approximation is accurate.

Plot (a), (b), (c) and (d) corresponds to setting $(tmax = 25, n = 50)$, $(tmax = 25, n = 100)$, $(tmax = 50, n = 25)$ and $(tmax = 50, n = 100)$ respectively. On each subplot, the lower diagonal plots contain pairwise parameter posterior points. The green dots represent posterior points from full MCMC integration via NUTS and the yellow points from enabled Laplace approximation of the random effects. The hollow red circles on the off-diagonal plots represent the pairwise means. The diagonal shows QQ-plot of posterior samples from Bayesian inference without (yellow dots) and with (green dots) Laplace approximation for that parameter including a 1:1 line in yellow. Even though the posterior points are densely packed, the overlap of the red circles with each technique shows seemingly good alignment of the two versions of the posterior, and this suggests that the Laplace approximation to the marginal likelihood where random effects are integrated out works well. Similar pair plots for Laplace approximation check with Prior1 can be found in Supporting Information.

4 Real-data case study

Having established the utility of our modeling approach and frequentist and Bayesian inference in the context of simulated data, we also applied the same statistical model to the analysis of a real great tit dataset of practical interest. The observed data were collected from a Dutch great tit (*Parus major*) population at the Hoge Veluwe National Park in the Netherlands ($52^{\circ}02'$ - $52^{\circ}07'$ N, $5^{\circ}51'$ - $5^{\circ}32'$ E). The recorded variables include the number of chicks, number of fledglings, mother ID, brood laying date and so on for each brood. Laying dates are presented as the number of days after March 31 (day 1=April 1, day 31=May 1). Similar to Reed et al. (2013), only the broods with one or more chicks were considered in our analysis due to the high proportion (15.7%) of zero-observations in the number of fledglings among the broods. The number of fledglings was taken as the fitness component and assumed to be Poisson distributed. The analyzed dataset consists of brood records breeding in 61 years from 1955 to 2015 and the sample size in a specific year ranges from 10 to 164 with an average of 81 across the study years. See Reed et al. (2013) for more details on the study population and fieldwork procedures.

The focus of this empirical study is to compare the computational efficiency of Bayesian inference with and without Laplace approximation and to check the accuracy of Laplace approximation. However, since the true structure of the model is unknown, we first conducted model selection under the frequentist framework and the candidate models considered are different from each other only in the model structure of stochastic α_t , θ_t and ω_t . The details of all the candidate models including the best model are given in Supporting Information. We then made Bayesian inference with the two different priors as in the simulation study using the selected model. For each prior distribution, we implemented `tmbstan` with and without Laplace approximation to check the accuracy of Laplace approximation.

Table 2 lists the reported estimates of model parameters from maximum likelihood (MLE) and Bayesian estimates with half-Cauchy (0, 10) prior (Prior1) and log-normal (1, 0.5) prior (Prior2). The best model indicates VAR(1) structure of α_t and θ_t and non-zero correlation $\hat{\rho}_{\alpha,\theta}$. The width of stabilizing fitness function turned to be constant over the study years implied by zero $\hat{\omega}_t$. Frequentist inference and Bayesian inference with Prior2 report close estimates for $\phi_{\theta,\theta}$ but the estimates with Prior2 show again less uncertainty for most of the estimates except for $\rho_{\alpha,\theta}$. In terms of $\log \sigma_\theta$, Bayesian inference with Prior1

Table 2: Frequentist and Bayesian estimates of parameters in the selected model with great tit dataset. The Bayesian estimates (in column Prior1 and Prior2) are obtained without Laplace approximation done by TMB.

parameter	MLE	Prior1	Prior2
μ_α	2(0.0369)	2(0.0491)	2(0.0379)
μ_θ	18.5(5.35)	18.8(7.12)	19.4(5.09)
μ_ω	3.88(0.055)	3.89(0.0563)	3.86(0.0522)
$\phi_{\alpha,\alpha}$	0.379(0.12)	0.458(0.13)	0.398(0.124)
$\phi_{\theta,\theta}$	0.48(0.112)	0.545(0.114)	0.477(0.102)
$\log\sigma_\alpha$	-1.72(0.14)	-1.63(0.152)	-1.76(0.126)
$\log\sigma_\theta$	3.07(0.137)	3.16(0.155)	2.98(0.125)
$\rho_{\alpha,\theta}$	-0.728(0.0825)	-0.715(0.0895)	-0.661(0.0987)

Table 3: Comparison of computational efficiency between Bayesian inference without (in the row "Full MCMC") and with Laplace approximation (in the row "Laplace approximation") for random effects for the great tit case study.

Model	Inference	Time(s)	min.ESS	Efficiency(ESS/t)
Prior 1	Full MCMC	1542.215	186.7651	0.1211019
	Laplace approximation	15491.85	1004.643	0.06484975
Prior 2	Full MCMC	1266.096	291.0717	0.229897
	Laplace approximation	7815.218	1111.257	0.1421914

reports the largest estimate and least certainty compared with the other two techniques. The close resemblance between estimates of $\log\sigma_\theta$ based on maximum likelihood and Bayesian inferences suggests that the data contains a good amount of information on $\log\sigma_\theta$ so that the maximum likelihood overwhelms the log-normal prior and dominates the posterior likelihood.

Table 3 shows computational efficiencies of Bayesian inference without and with Laplace approximation. It turns out that the computational efficiency with Laplace approximation is approximately half of that without Laplace approximation in both models with Prior1 and Prior2.

Similar to Fig. 4, Fig. 5 and Fig. 6 display pair plots of posterior samples to check the accuracy of Laplace approximation with Prior1 and Prior2 respectively. Both the figures seemingly suggest a good mix of posterior samples with and without Laplace approximation for all the parameters in the selected model, indicating that the Laplace approximation assumption is met.

5 Conclusions and extensions

In this study, we have investigated frequentist inference and Bayesian inference with two different priors. The inferences were implemented with a state-space model estimating temporal fluctuating selection and with simulated biological data under four different simulation settings. A state-of-the-art R package (tmbstan) for fast fitting statistical models was used for Bayesian inference with Laplace approximation turning on or off. The simulation studies show that the choice of prior can have an important impact on the geometric

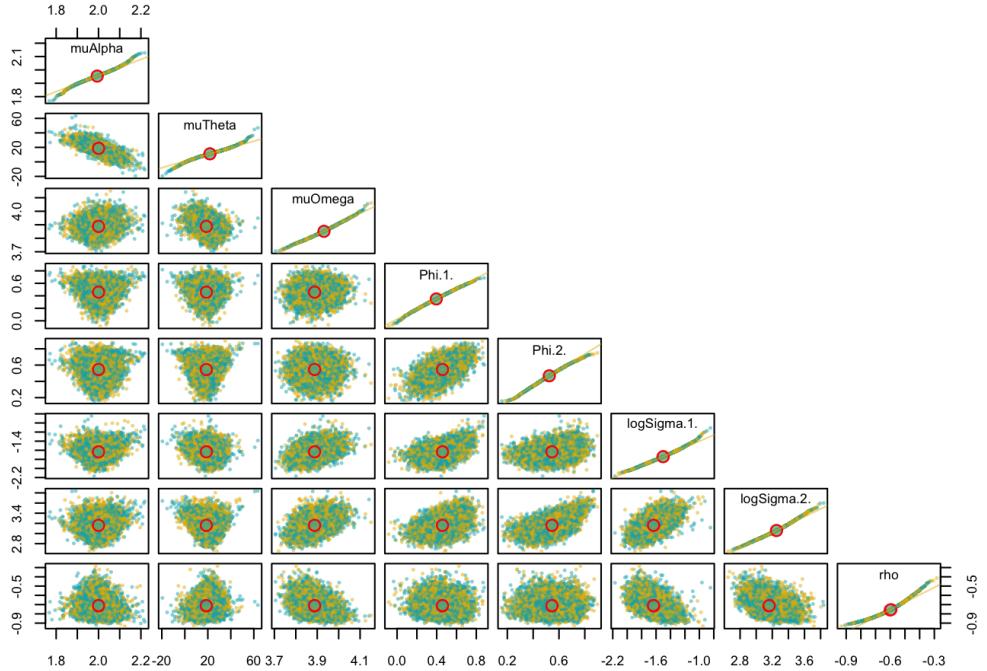


Figure 5: Pair plots of posterior samples for Laplace approximation test for the great tit case study with Prior1. The random effects in the great tit TMB model can be integrated with two techniques: (1) full MCMC integration via NUTS and (2) Laplace approximation. To check the accuracy of Laplace approximation to the posterior likelihood density, the posterior samples for all the fixed effects in the model without (yellow dots) and with Laplace approximation (green dots) are shown pair-wisely on the same plot. Columns and rows on the lower diagonal correspond to pair-wise parameters, with the diagonal showing QQ-plot of posterior samples from Bayesian inference without (yellow dots) and with (green dots) Laplace approximation for that parameter including a 1:1 line in yellow. The large red circles of the off-diagonal plots represent the pairwise means. On each off-diagonal plot, there are 4000 yellow dots corresponding to 1000 samples retained from each of four chains without Laplace approximation, so as the green dots with Laplace approximation. Posterior rows were randomized to prevent consistent overplotting of one integration technique. Overlaps in the two colored dots suggest the Laplace approximation assumption is met.

shape of the posterior distributions of the model parameters and a non-informative prior (in this study uniform prior and inverse-gamma prior on the scale parameter) may lead to unstable inference since the Markov chains may not converge or get stuck in part of the ridge of posterior. With unobserved states following a VAR(1) process, we also found that the autoregressive parameters and the scale parameters in the variance-covariance matrix of the states are difficult and challenging to be estimated accurately. The increased sample size at each time point does not necessarily provide more information for

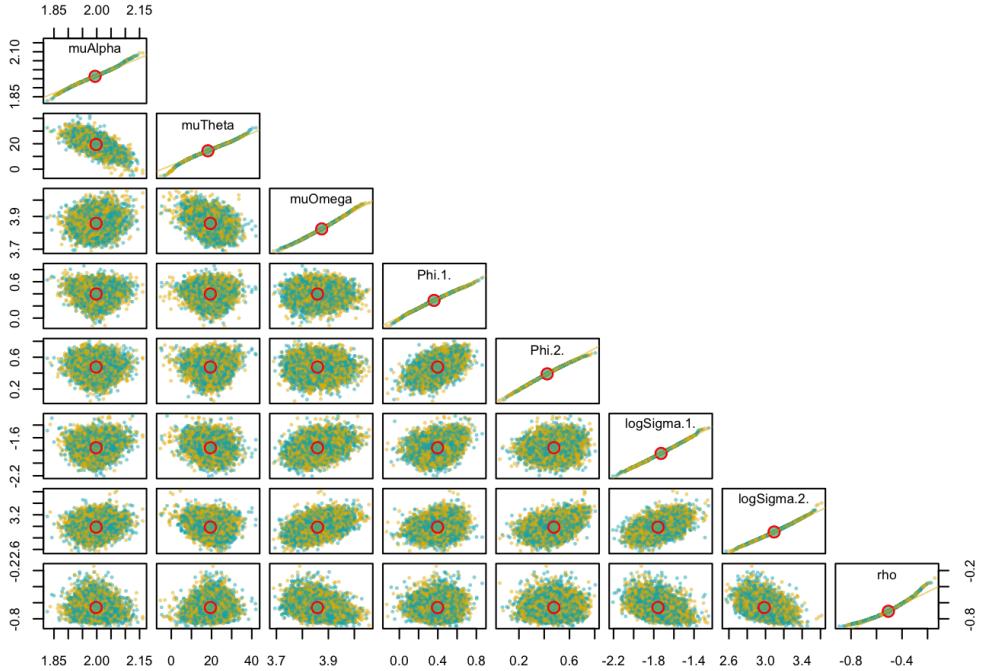


Figure 6: Pair plots of posterior samples for Laplace approximation test for the great tit case study with Prior2. The random effects in the great tit TMB model can be integrated with two techniques: (1) full MCMC integration via NUTS and (2) Laplace approximation. To check the accuracy of Laplace approximation to the posterior likelihood density, the posterior samples for all the fixed effects in the model without (yellow dots) and with Laplace approximation (green dots) are shown pair-wisely on the same plot. Columns and rows on the lower diagonal correspond to pair-wise parameters, with the diagonal showing QQ-plot of posterior samples from Bayesian inference without (yellow dots) and with (green dots) Laplace approximation for that parameter including a 1:1 line in yellow. The large red circles of the off-diagonal plots represent the pairwise means. On each off-diagonal plot, 4000 yellow dots correspond to 1000 samples retained from each of four chains without Laplace approximation, so as the green dots with Laplace approximation. Posterior rows were randomized to prevent consistent overplotting of one integration technique. Overlaps in the two colored dots suggest the Laplace approximation assumption is met.

the transition parameters and scale parameters. Only more time points in the data could make the likelihood dominate the posterior likelihood and thus lead to better estimates of these parameters. Half-Cauchy prior on the scale parameter leads to less stable inference than log-normal prior indicated by the number of divergent transitions in the Markov Chains. Laplace approximation for the random effects turns out to be accurate suggested by the pair plots of the posterior samples with and without Laplace approximation for both the simulation studies and the great tit case study. Turning on Laplace approxi-

mation in `tmbstan` would probably reduce computational efficiency but it is worth trying when there is a good amount of data, in which case the Laplace approximation is more likely to be accurate and also potentially improve the computational efficiency of MCMC.

In our study, we used arbitrary prior distributions, however, the prior information can be obtained from different sources. For example, in our great tit case study, the timing and width of the caterpillar peak can provide a clue for the time window of optimal laying dates, thus the information can be used to decide the prior for the scale parameter of the optimal laying dates. Prior information can also be generated from previous studies on the same species and more general ecological knowledge coming from other related species (Tufto et al., 2000).

We conducted simulation studies with only AR(1) process of the optimal laying dates, but the model is formulated and coded in a way that can be effortlessly extended to order-1 vector autoregression (VAR(1)). It can be widely used for modeling ecological processes where auto-correlation and cross-correlation in the processes arise due to shared environmental variables at either temporal or spatial scale. We expect more ecologists to adopt these two new estimation methods, TMB, and `tmbstan`, given its flexibility in either frequentist or Bayesian inference for a wide range of models, including the models where the unobserved ecological processes are treated as latent variables and assumed to be VAR processes. However, the drawback of Bayesian VAR (BVAR) methods is that it usually requires estimation of a large number of parameters and thus the over-parameterization might lead to unstable inference and inaccurate out-of-sample forecasts. Some shrinkage methods (Sims & Zha, 1998; Koop et al., 2010; Giannone et al., 2015; Sørbye & Rue, 2017, for example) were thereby developed, in which Bayesian priors provide a logical and consistent method of imposing parameter restrictions that can be potentially applied to ecological data cases.

References

- Albertsen, C. M., Nielsen, A., & Thygesen, U. H. (2016). Choosing the observational likelihood in state-space stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(5), 779–789.
- Auger-Méthé, M., Albertsen, C. M., Jonsen, I. D., Derocher, A. E., Lidgard, D. C., Studholme, K. R., ... Flemming, J. M. (2017). Spatiotemporal modelling of marine movement data using Template Model Builder (TMB). *Marine Ecology Progress Series*, 565, 237–249.
- Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79, 30.
- Cadigan, N. G. (2015). A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(2), 296–308.
- Cao, Y., Visser, M. E., & Tufto, J. (2019). A time-series model for estimating temporal variation in phenotypic selection on laying dates in a dutch great tit population. *Methods in Ecology and Evolution*, 10(9), 1401–1411.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).
- Chevin, L.-M., Visser, M. E., & Tufto, J. (2015). Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. *Evolution*, 69(9), 2319–2332.
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., & Riebler, A. (2019). Intuitive principle-based priors for attributing variance in additive model structures. *arXiv preprint arXiv:1902.00242*.
- Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5), 530–543.
- Gelman, A., et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515–534.
- Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2), 436–451.
- Grant, P. R., & Grant, B. R. (2002). Unpredictable evolution in a 30-year study of Darwin's finches. *science*, 296(5568), 707–711.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.
- Jeffreys, H., & Jeffreys, H. (1961). *Theory of probability (3rd edn)*. Oxford.
- Karakani, H. M., van Niekerk, J., & van Staden, P. (2016). Bayesian analysis of AR (1) model. *arXiv preprint arXiv:1611.08747*.
- Koop, G., Korobilis, D., et al. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, 3(4), 267–358.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5), 1–21.

- Lei, G., Boys, R., Gillespie, C., Greenall, A., & Wilkinson, D. (2011). Bayesian inference for sparse VAR (1) models, with application to time course microarray data. *Journal of Biometrics and Biostatistics*.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4), 325–337.
- Monnahan, C. C., & Kristensen, K. (2018). No-U-turn sampling for fast bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PloS one*, 13(5), e0197954.
- Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3), 339–348.
- Neal, R. M., et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11), 2.
- Reed, T. E., Jenouvrier, S., & Visser, M. E. (2013). Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. *Journal of Animal Ecology*, 82(1), 131–144.
- Schlaifer, R., & Raiffa, H. (1961). *Applied statistical decision theory*. Wiley Cambridge.
- Sims, C. A., & Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 949–968.
- Sørbye, S. H., & Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6), 923–935.
- Stan Development Team. (2018a). Rstan: the R interface to stan. *R package version 2.17.3*. <http://mc-stan.org>.
- Stan Development Team. (2018b). Stan modeling language users guide and reference manual. *Version 2.18.0*. <http://mc-stan.org>.
- Tufto, J., Lande, R., Ringsby, T.-H., Engen, S., Sæther, B.-E., Walla, T. R., & DeVries, P. J. (2012). Estimating Brownian motion dispersal rate, longevity and population density from spatially explicit mark-recapture data on tropical butterflies. *Journal of Animal Ecology*, 81(4), 756–769.
- Tufto, J., Sæther, B.-E., Engen, S., Arcese, P., Jerstad, K., Røstad, O. W., & Smith, J. N. (2000). Bayesian meta-analysis of demographic parameters in three small, temperate passernines. *Oikos*, 88(2), 273–281.
- Vedder, O., Bouwhuis, S., & Sheldon, B. C. (2013). Quantitative assessment of the importance of phenotypic plasticity in adaptation to climate change in wild bird populations. *PLoS Biology*, 11(7), e1001605.
- Wei, W. (2006). *Time series analysis: Univariate and multivariate methods* (2nd ed.).
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

Supporting Information (SI) for

Bayesian inference with `tmbstan` for a state-space model with VAR(1) state equation

1 Supplementary results of simulation studies

Similar to Table 1 in the main text, we here show the frequentist and Bayesian estimates of the same parameters but with different true values of $\phi_{\theta,\theta}$. Table S1 and Table S2 list the estimates of parameters under different simulation settings with $\phi_{\theta,\theta} = 0.1$ and 0.7 respectively. From these two tables, we find generally similar patterns to the table of estimates in the main text. For example, dataset with more time points ($tmax = 50, n = 100$) leads to more accurate estimates compared with the dataset with shorter time series ($tmax = 25, n = 100$). Increasing the sample size at each time point improves neither the accuracy nor the certainty of the estimates for the parameters of interest, only a bigger sample size is required for this purpose.

In the main text, we only present the pair plots of posterior samples for Laplace approximation check with Prior2. We here supplement the pair plots (Fig. S1) with Prior1 under the four different sample size settings. Fig. S1 also suggests accurate Laplace approximation indicated by the good mix of posterior samples. To further validate this conclusion, we visually inspect the accuracy of Laplace approximation by plotting bivariate contour plots of posterior samples from the Bayesian model with and without Laplace approximation on the same figure, as shown in Fig. S2. Only the joint posterior distribution ($\phi_{\theta,\theta}$ and $\log(\sigma_\theta)$) is considered and other parameters are ignored for simplifying the analysis. The overlap of contours with (yellow) and without (green) Laplace approximation for the random effects suggests again that the Laplace approximation in these cases is accurate.

Table S1: Frequentist and Bayesian estimates from the model with AR(1) θ_t , autocorrelation in θ_t $\phi_{\theta,\theta} = 0.1$, and different sample sizes.

$\phi_{\theta,\theta} = 0.1, tmax = 25, n = 50$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	1	0
μ_α	2	2.006(0.016)	2.005(0.016)	2.006(0.016)
μ_θ	20	19.6(6.4)	19.3(9.3)	19.8(6.1)
μ_ω	3.5	3.475(0.030)	3.479(0.030)	3.472(0.030)
$\phi_{\theta,\theta}$	0.1	0.26(0.19)	0.34(0.22)	0.25(0.16)
$\log\sigma_\theta$	2.996	3.21(0.16)	3.34(0.20)	3.11(0.14)
$\phi_{\theta,\theta} = 0.1, tmax = 25, n = 100$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	1	0
μ_α	2	1.996(0.010)	1.996(0.010)	1.997(0.010)
μ_θ	20	17.1(3.7)	16.4(5.0)	17.0(3.8)
μ_ω	3.5	3.493(0.021)	3.494(0.022)	3.491(0.022)
$\phi_{\theta,\theta}$	0.1	0.07(0.21)	0.15(0.24)	0.10(0.18)
$\log\sigma_\theta$	2.996	2.85(0.15)	2.95(0.18)	2.78(0.13)
$\phi_{\theta,\theta} = 0.1, tmax = 50, n = 25$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	0	0
μ_α	2	1.977(0.015)	1.977(0.015)	1.976(0.015)
μ_θ	20	19.8(2.7)	19.7(3.1)	19.9(2.8)
μ_ω	3.5	3.529(0.033)	3.535(0.033)	3.525(0.033)
$\phi_{\theta,\theta}$	0.1	0.04(0.15)	0.07(0.17)	0.06(0.14)
$\log\sigma_\theta$	2.996	2.88(0.12)	2.93(0.12)	2.83(0.12)
$\phi_{\theta,\theta} = 0.1, tmax = 50, n = 100$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	0	0
μ_α	2	1.9858(0.0076)	1.9857(0.0078)	1.9856(0.0077)
μ_θ	20	20.3(2.8)	20.3(2.9)	20.3(2.9)
μ_ω	3.5	3.515(0.015)	3.515(0.016)	3.513(0.015)
$\phi_{\theta,\theta}$	0.1	0.09(0.14)	0.12(0.16)	0.11(0.14)
$\log\sigma_\theta$	2.996	2.89(0.10)	2.93(0.11)	2.86(0.10)

Table S2: Frequentist and Bayesian estimates from the model with AR(1) θ_t , autocorrelation in θ_t $\phi_{\theta,\theta} = 0.7$, and different sample sizes.

$\phi_{\theta,\theta} = 0.7, tmax = 25, n = 50$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	1	0
μ_α	2	2.012(0.015)	2.012(0.015)	2.011(0.014)
μ_θ	20	15.7(4.8)	15.0(8.7)	16.0(5.0)
μ_ω	3.5	3.483(0.031)	3.486(0.031)	3.480(0.031)
$\phi_{\theta,\theta}$	0.7	0.45(0.18)	0.55(0.19)	0.43(0.16)
$\log\sigma_\theta$	2.996	2.72(0.18)	2.89(0.27)	2.65(0.16)
$\phi_{\theta,\theta} = 0.7, tmax = 25, n = 100$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	NA	0
μ_α	2	1.987(0.011)	1.980(0.014)	1.986(0.011)
μ_θ	20	18.3(9.7)	20(18)	18.4(8.1)
μ_ω	3.5	3.539(0.022)	3.566(0.049)	3.537(0.022)
$\phi_{\theta,\theta}$	0.7	0.70(0.13)	0.60(0.35)	0.63(0.11)
$\log\sigma_\theta$	2.996	3.10(0.23)	3.36(0.29)	2.95(0.16)
$\phi_{\theta,\theta} = 0.7, tmax = 50, n = 25$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	1	0
μ_α	2	2.021(0.016)	2.021(0.016)	2.021(0.016)
μ_θ	20	19.3(8.9)	20(14)	19.7(7.8)
μ_ω	3.5	3.488(0.031)	3.490(0.030)	3.482(0.030)
$\phi_{\theta,\theta}$	0.7	0.739(0.094)	0.781(0.091)	0.692(0.081)
$\log\sigma_\theta$	2.996	3.24(0.18)	3.39(0.25)	3.13(0.14)
$\phi_{\theta,\theta} = 0.7, tmax = 50, n = 100$				
Parameters	True value	MLE	Prior1	Prior2
no. divergent transitions	NA	NA	1	0
μ_α	2	1.9899(0.0076)	1.9899(0.0076)	1.9896(0.0075)
μ_θ	20	21.1(6.2)	20(12)	21.6(5.5)
μ_ω	3.5	3.511(0.015)	3.511(0.015)	3.510(0.015)
$\phi_{\theta,\theta}$	0.7	0.71(0.10)	0.76(0.10)	0.667(0.086)
$\log\sigma_\theta$	2.996	2.93(0.17)	3.09(0.27)	2.84(0.14)

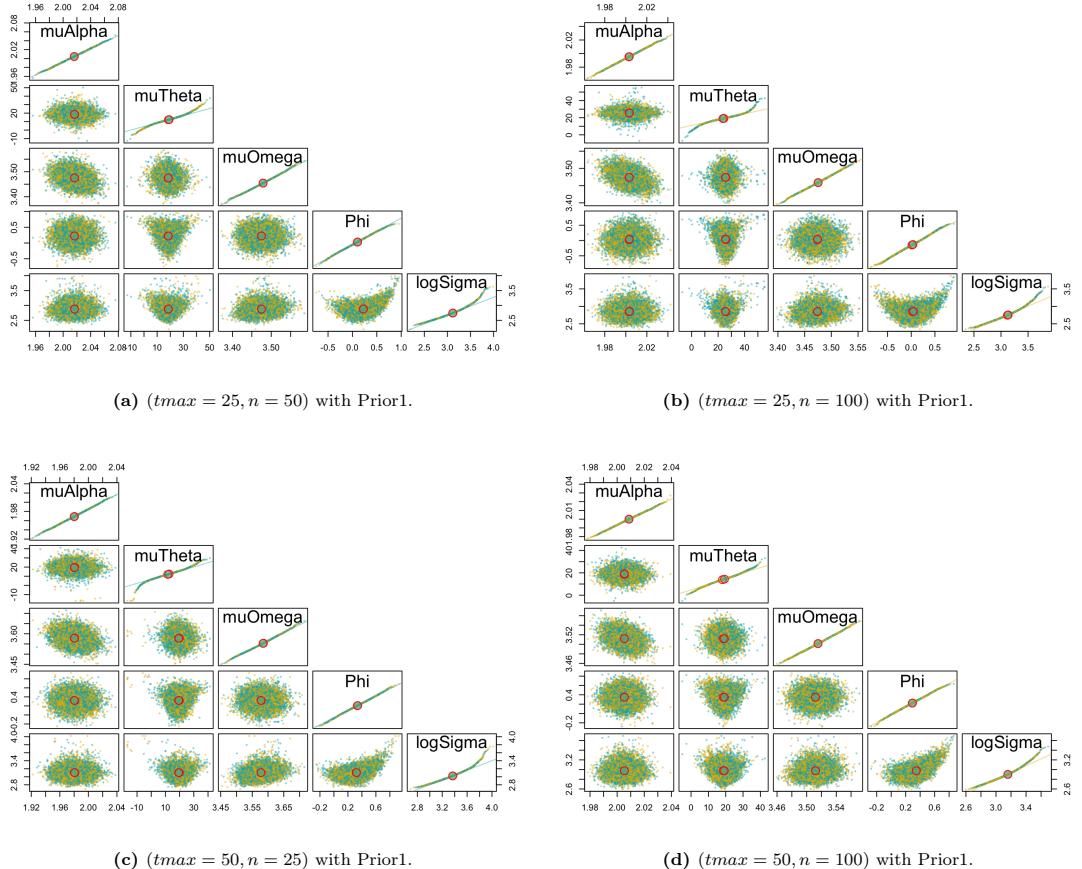
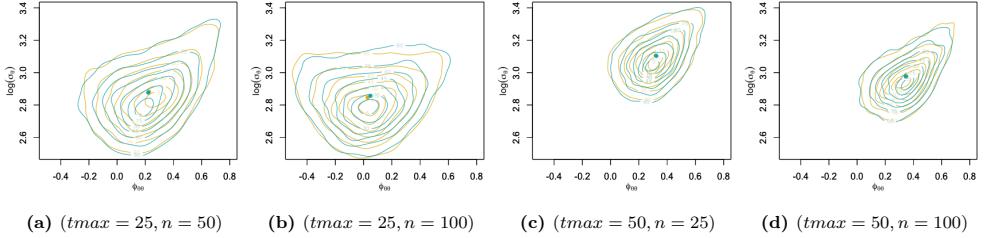


Figure S1: Pair plots of posterior samples for Laplace approximation check for one realization of the simulation with prior1. The four plots (a), (b), (c), and (d) correspond to the four schemes of simulation. The random effects in the TMB model can be integrated with two techniques: (1) full MCMC integration via NUTS and (2) Laplace approximation. To check the accuracy of Laplace approximation to the posterior likelihood density, the posterior samples for all the fixed effects in the model without (yellow dots) and with Laplace approximation (green dots) are shown pair-wisely on the same plot. Columns and rows on the lower diagonal correspond to pair-wise parameters, with the diagonal showing QQ-plot of posterior samples from Bayesian inference without (yellow dots) and with (green dots) for that parameter including a 1:1 line in yellow. The large red circles of the off-diagonal plots represent the pairwise means. On each off-diagonal plot, there are 4000 yellow dots corresponding to 1000 samples retained from each of four chains without Laplace approximation, so as to the green dots with Laplace approximation. Posterior rows were randomized to prevent consistent overplotting of one integration technique. Overlaps in the two colored dots suggest the Laplace approximation assumption is met.

Prior1



Prior2

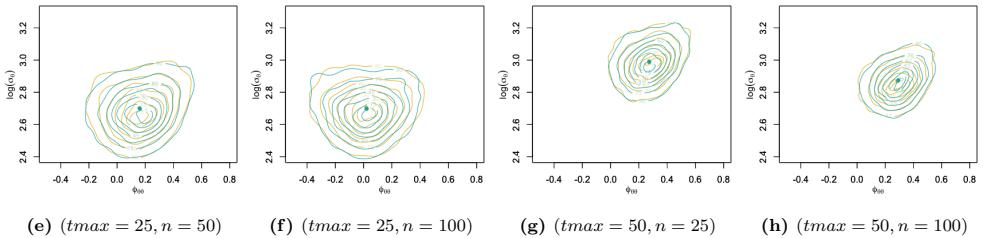


Figure S2: Bivariate contour plots of posterior samples of $\phi_{\theta,\theta}$ and $\log(\sigma_\theta)$ from one realization of the simulation with Prior1 (the first row) and Prior2 (the second row) for Laplace approximation check. The posterior samples data used are the same as that in Fig. S1 and Figure 4 in the main text. The yellow contours indicate the joint posterior distribution of $(\phi_{\theta,\theta}, \log(\sigma_\theta))$ from the estimation technique full MCMC integration via NUTS, and the green contours correspond to the technique that Laplace approximation is used. The yellow and green dots in each plot represent the mean of the bivariate posterior samples in each setting respectively.

Table S3: Model selection for the real data case study. The table lists all the candidate models fitted with the great tit data. Model 7 is selected as the best model due to the smallest AIC value. Column Δp and ΔAIC lists the difference between the selected model and the corresponding candidate model in the number of parameters and reported AIC value respectively. The rightmost column describes the candidate models. The elements in matrix Φ and vector ρ are set to 0 if not otherwise specified.

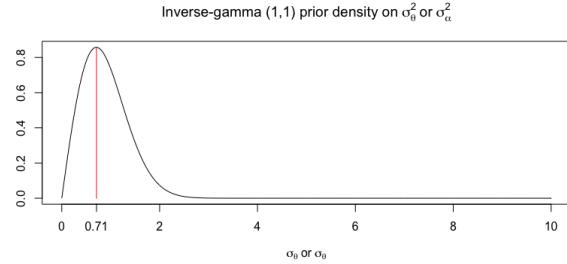
Model	Δp	ΔAIC	Description
1	-5	821.51	$\alpha_t = \theta_t = \omega_t = 0$
2	-4	295.07	$\theta_t = \omega_t = 0$, random α_t
3	-3	34.81	$\omega_t = 0$, random α_t and θ_t
4	-2	176.51	random α_t , θ_t and ω_t
5	-4	265.32	$\alpha_t = \omega_t = 0$, random θ_t
6	2	2.52	$\omega_t = 0$, VAR(1) α_t and θ_t : $\phi_{\alpha,\alpha} \neq \phi_{\theta,\theta} \neq \phi_{\alpha,\theta} \neq \phi_{\theta,\alpha} \neq 0$, $\rho_{\alpha,\theta} \neq 0$
7 (best model)	0	0	$\omega_t = 0$, AR(1) α_t and AR(1) θ_t : $\phi_{\alpha,\alpha} \neq \phi_{\theta,\theta} \neq 0$
8	1	1.92	$\omega_t = 0$, VAR(1) α_t and θ_t : $\phi_{\alpha,\alpha} \neq \phi_{\theta,\theta} \neq \phi_{\theta,\alpha} \neq 0$, $\rho_{\alpha,\theta} \neq 0$
9	1	1.21	$\omega_t = 0$, VAR(1) α_t and θ_t : $\phi_{\alpha,\alpha} \neq \phi_{\theta,\theta} \neq \phi_{\alpha,\theta} \neq 0$, $\rho_{\alpha,\theta} \neq 0$
10	-1	12.93	$\omega_t = 0$, random θ_t , AR(1) α_t : $\phi_{\alpha,\alpha} \neq 0$
11	-1	6.7	$\omega_t = 0$, random α_t , AR(1) θ_t : $\phi_{\theta,\theta} \neq 0$

2 Supplementary info on real data case study

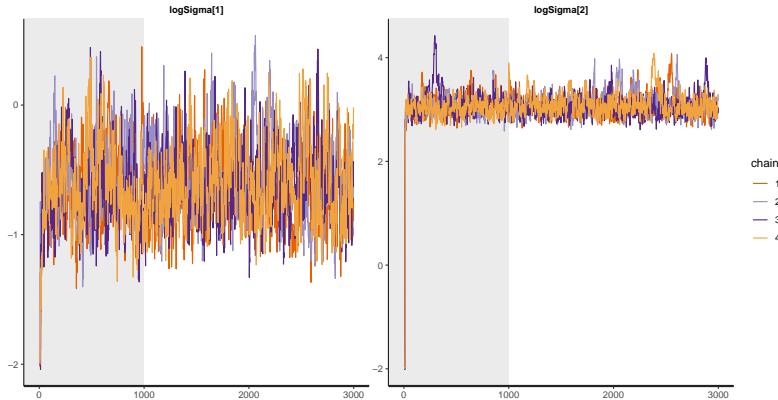
Beside half-Cauchy and lognormal priors for the scale parameters of the great tit model as shown in the main text, we also tested inverse-gamma (1, 1) prior for the scale parameter σ_α^2 and σ_θ^2 . To visualize MCMC diagnostics we show trace plots for the two scale parameters along with the prior densities in Fig. S3. The solid black line in plot (a) shows prior density function of σ_α (or σ_θ) given a Inverse-gamma (1, 1) prior density on σ_α^2 (or σ_θ^2). The details on density function transformation are omitted here. The solid red line indicates the density mode. The prior density mode of σ_α at 0.71 translates to density mode of $\log\sigma_\alpha$ at -0.34. However, the left trace plot in plot (b) for $\log\sigma_\alpha$ implies that the posterior likelihood is dominated by the prior so that the sampler gets trapped in the subspace of the parameter, which is a space near -0.34, while the true posterior density mode locates around -1.7.

As mentioned in the main text, the great tit model implemented with Bayesian inference was selected in the frequentist framework with model selection procedure. Table S3 lists all the candidate models fitted with the great tit data. Model 7 is selected as the best model due to the smallest AIC value reported. Colomun Δp and ΔAIC lists the difference between the selected model and the corresponding candidate model in the number of parameters and reported AIC value respectively. The rightmost column describes the candidate models.

We also plot the contours of posterior samples with and without Laplace approximation for a subset



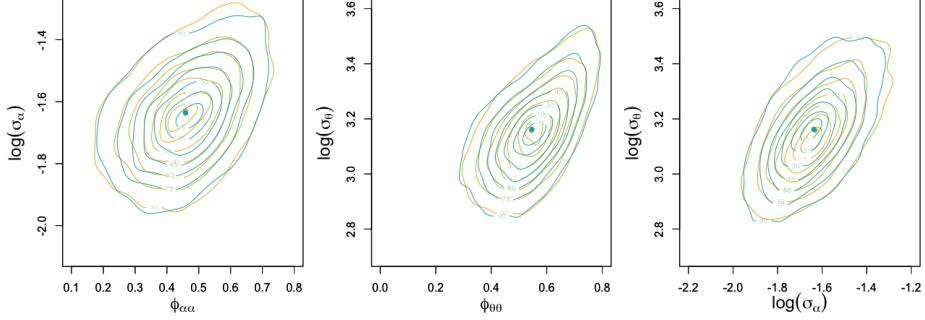
(a) Prior density function on σ_α or σ_θ given Inverse-gamma (1, 1) prior on σ_α^2 or σ_θ^2 respectively.



(b) Trace plots for $\log\sigma_\alpha$ (left) and for $\log\sigma_\theta$ (right).

Figure S3: A prior density and trace plots for the great tit case study. In plot (a), the solid curve indicates an equivalence of the density to inverse-gamma (1, 1) prior on σ_α^2 or σ_θ^2 , the equivalent density on σ_α or σ_θ is calculated with rules of density function transformation, which is omitted here. The red solid line indicates the density mode. Plot (b) shows trace plots with the inverse-gamma (1,1) priors for parameter σ_α^2 (left) and σ_θ^2 (right) respectively. The grey areas indicate warm-up iterations.

(a) Bivariate contour plots with Prior1.



(b) Bivariate contour plots with Prior2.

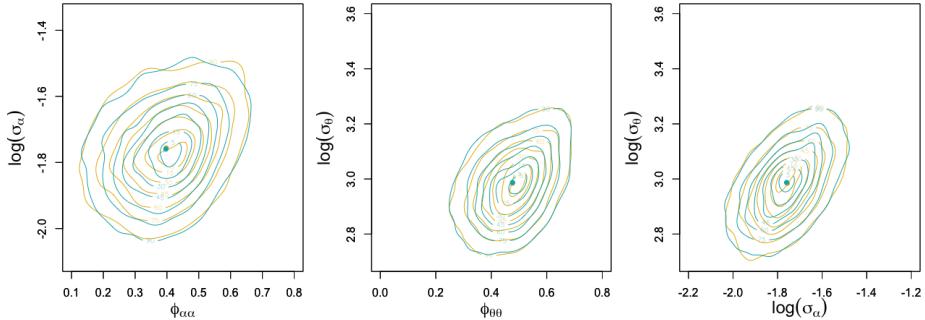


Figure S4: Bivariate contour plots of posterior samples of a subset of the parameters in the selected great tit model for Laplace approximation check. The posterior samples used here are the same as that in Figure 5 and Figure 6 in the main text. The plots in row (a) correspond to the Bayesian model with Prior1, and in row (b) they are with Prior2. Similar to Fig. S2, the yellow contours indicate the joint posterior distribution of the parameters from the estimation technique full MCMC integration via NUTS, and the green contours correspond to the technique that Laplace approximation is used. The yellow and green dots in each plot again represent the mean of the bivariate posterior samples in each plot respectively. Only a subset of the parameters is considered for simplification.

of parameters in the great tit model on the same graph (Fig. S4), to get a clearer visualization of the posteriors' distribution. The first and second row of the contour plots corresponds to the Bayesian great tit model with Prior1 and Prior2 respectively. The round dots on the plots are the mean of posterior samples for each estimation technique. The good amount of overlap of the yellow contours, dots (without Laplace approximation), and green contours, dots (with Laplace approximation) again suggests a good accuracy of Laplace approximation.