

Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection

Luis-Miguel Chevin¹, Marcel E. Visser², and Jarle Tufto³

July 17, 2015

¹CEFE-CNRS, UMR 5175, 1919 route de Mende, 34293 Montpellier 05, France

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

³Centre for Biodiversity Dynamics/Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Abstract

Despite considerable interest in temporal and spatial variation of phenotypic selection, very few methods allow quantifying this variation while correctly accounting for the error variance of each individual estimate. Furthermore, the available methods do not estimate the autocorrelation of phenotypic selection, which is a major determinant of eco-evolutionary dynamics in changing environments. We introduce a new method for measuring variable phenotypic selection using random regression. We rely on model selection to assess the support for stabilizing selection, and for a moving optimum that may include a trend plus (possibly autocorrelated) fluctuations. The environmental sensitivity of selection also can be estimated by including an environmental covariate. After testing our method on extensive simulations, we apply it to breeding time in a great tit population in the Netherlands. Our analysis finds support for an optimum that is well predicted by spring temperature, and occurs about 33 days before a peak in food biomass, consistent with what is known from the biology of this species. We also detect autocorrelated fluctuations in the optimum, beyond those caused by temperature and the food peak. Because our approach directly estimates parameters that appear in theoretical models, it should be particularly useful for predicting eco-evolutionary responses to environmental change.

Keywords. Fluctuating selection; Gaussian fitness function; Generalized linear mixed models; Poisson regression.

Introduction

The extent to which natural selection changes in time and space has profound implications for the dynamics of adaptation, from the maintenance of polymorphism (reviewed by [Felsenstein, 1976](#); [Hedrick et al., 1976](#); [Hedrick, 2006](#); [Bell, 2010](#)) to the evolution of bet hedging and/or phenotypic plasticity ([Gillespie, 1973, 1974](#); [Bull, 1987](#); [Svardal et al., 2011](#); [Via and Lande, 1985](#); [Gavrilets and Scheiner, 1993](#); [Lande, 2009](#); [Tufto, 2015](#)). Apparent temporal variation in selection is a common feature of several classic examples of natural selection in the wild (reviewed in [Bell, 2010](#); [Calsbeek et al., 2012](#); [Svensson and Calsbeek, 2012](#)), from beak shape in Darwin’s finches ([Grant and Grant, 2002](#)) to banding patterns in *Cepea* snails ([Cain et al., 1990](#)), or spine number in three-spined sticklebacks ([Reimchen and Nosil, 2002](#)). Local adaptation also is pervasive ([Hereford, 2009](#)), indicating that natural selection often varies in space too. But despite a long-lasting interest in variable selection, the latter is often reported mostly as a qualitative pattern, and there are few quantitative measurements of the magnitude of variation in selection at the phenotypic level (but see [Engen et al., 2012](#); [Calsbeek et al., 2012](#)). In particular, the few available methods do not treat phenotypic selection across time or space as an actual stochastic process, but rather as a simple random variable, and thus provide no measure of its autocorrelation.

An important aspect of variation in phenotypic selection is the relationship between phenotypic selection and the environment. Conceptually, for a phenotypic trait to be involved in adaptation to particular axes of the ecological niche (*sensu* Hutchinson’s multidimensional niche, [Holt \(2009\)](#)), selection on that trait needs to change with environmental variables that define these axes. The causes of natural selection (or selective agents) can thus be revealed by demonstrating a covariance between measurements of selection and environmental variables ([Wade and Kalisz, 1990](#)), but a recent review highlighted that this has seldom been performed so far ([MacColl, 2011](#)), except for a few notable examples ([Ahola et al., 2012](#); [Kelly, 1992](#); [Brown et al., 2013](#); [Reed et al., 2013b](#); [Visser et al., 2015](#)). Two obvious reasons are the difficulty in identifying (or measuring)

relevant environmental variables, and having enough datapoints in time or space to estimate the selection-environment covariance. Another reason is that the importance of the covariance between environment and phenotypic selection has perhaps not been fully appreciated until recently, at least outside of the theoretical literature. This situation is changing, however, and the environmental sensitivity of selection (change in an optimum phenotype with the environment), a key component of theories on phenotypic adaptation and phenotypic plasticity, is increasingly perceived as an important determinant of the interplay between genetic evolution, phenotypic plasticity, and population growth in a changing environment (Chevin and Lande, 2010; Gienapp et al., 2013; Vedder et al., 2013; Michel et al., 2014). A third impediment is the lack of a robust and general method for estimating the variance of phenotypic selection, and its covariance with environmental variables. One of our aims is to provide such a method.

Most measurements of phenotypic selection in a given generation rely on methods stemming from the Lande and Arnold (1983) approach, which is based on classic linear regression. Extensions of this method allow treating more realistic distributions of fitness residuals and trait-fitness relationships by use of generalized linear models. For instance, Janzen and Stern (1998) proposed using logistic regression (with a logit link function and Bernoulli or Binomial responses) for analysis of survival selection. More recently, aster modeling has been developed to decompose fitness into components with known distributions, and fit these components jointly to infer selection (Shaw and Geyer, 2010). The Lande-Arnold approach also has been extended recently by Morrissey and Sakrejda (2013) to allow for measurements of selection gradients from arbitrary fitness functions estimated by spline-based methods (Schluter, 1988), rather than by the original quadratic approximation.

In contrast, there are fewer developments on measuring the *variability* of phenotypic selection. Even when phenotypic selection is shown to vary significantly through space or time, the magnitude of this variation often is not quantified (Ahola et al., 2012; Kelly, 1992; Brown et al., 2013). When measurements of phenotypes and fitnesses are available for different times (or locations), the most common procedure has been to estimate each

selection gradient separately, and then compute the variance of selection gradients from the variance of each estimates (as reviewed by [Siepielski et al., 2009](#)). This is however problematic, as it conflates sampling variance with the actual variance of the process, as highlighted by [Morrissey and Hadfield \(2012\)](#) (see also [Siepielski et al. \(2013\)](#)). A method has recently been proposed to estimate fluctuating selection in age-structured populations ([Engen et al., 2012](#)), which directly estimates the variance of selection gradients by maximum likelihood from the phenotypes and fitnesses in all years, correctly accounting for measurement error. However, this method assumes that selection gradients are independent and identically distributed, and therefore does not allow for autocorrelation (the same holds for a non-parametric approach proposed by [Calsbeek et al. \(2012\)](#)). Yet theory has shown that the temporal autocorrelation in phenotypic selection is at least as important as its variance for the long-term evolution and demography of a population ([Charlesworth, 1993](#); [Lande and Shannon, 1996](#); [Bürger and Gimelfarb, 2002](#); [Chevin, 2013](#); [Tufto, 2015](#); [Zhang, 2012](#)), so it is necessary to treat phenotypic selection explicitly as a time series.

One of the most appealing properties of the directional selection gradient is that, in a given generation, it directly relates to the quantity that predicts evolutionary change of the mean phenotype for a normally distributed trait, namely the slope of the adaptive surface relating (log) mean fitness to the mean trait (hence the name gradient, [Lande \(1976\)](#); [Lande and Arnold \(1983\)](#)). But in a temporally variable environment, although the variance of directional gradients can be used to predict the magnitude and variance of responses to selection (conditional on the additive genetic variance), it cannot reveal whether the genetic response to selection in a given generation is likely to be increase fitness in the next (as discussed in [Chevin \(2013\)](#)), which will often be of interest. This more complete description of the evolution of quantitative traits in fluctuating environments is afforded by models of a moving optimum, where fitness is maximized at an intermediate phenotypic value that changes with the environment ([Charlesworth, 1993](#); [Lande and Shannon, 1996](#); [Bürger and Gimelfarb, 2002](#); [Chevin, 2013](#); [Tufto, 2015](#)). In such models, directional selection is caused by a mismatch between the optimal and mean phenotype,

and the long-term evolution (and demography) of the population depends on patterns of movements of the optimum, which are not fully captured by temporal change in directional selection gradients (Chevin and Haller, 2014). Furthermore, fluctuating selection is likely to be a strong driver of the environmental stochasticity of population growth, but computing the contribution of fluctuating selection on some measured trait(s) to the overall observed fluctuations in population size requires more information than is provided by the temporal distribution of selection gradient (which only describes the local slope of the fitness function). Hence, if we are to estimate parameters of fluctuating phenotypic selection that can be related to theoretical predictions on expected fitness and population growth in a random environment, we should focus as much as possible on changes in the actual (absolute) fitness surface, rather than just on selection gradients. If the fitness function includes an optimum phenotype, then the location, width and height of the fitness peak should be estimated.

Here we introduce a method to estimate variation in phenotypic selection using measurements of traits and fitness (or components thereof) across time (or space). This method responds to the requirements detailed above, as well as several other important ones. First, fitness is treated as an expected property of individuals, rather than a realized number of offspring, for instance. This corresponds to the propensity definition of fitness, as generally accepted in population genetics (Mills and Beatty, 1979), and implemented in analysis of phenotypic selection for instance by Shaw and Geyer (2010). Considering fitness or its components as parameters of distributions of random variables (survival probability, expected number of offspring), rather than realizations of the variables themselves, allows separating the actual selection gradient/differential from the random component of phenotypic change caused by demographic stochasticity (Engen and Sæther, 2014). The latter contributes to random genetic drift rather than change caused by natural selection (Rice, 2004, pp. 181-187). Second, we use phenotypes and fitnesses from all years (or locations) to estimate variation in selection using random regression, rather than estimating selection separately within each year (or location). This allows a more accurate estimation of parameters, ruling out for instance apparent fluctuation

tuating selection caused by estimation error across years (Siepielski et al., 2009). This is especially relevant for quadratic terms describing stabilizing (if negative) or disruptive (if positive) selection. These two forms of selection have important qualitative differences in terms of the evolutionary predictions they generate (Lande, 1976; Bulmer, 1974; Bürger and Gimelfarb, 2002; Dieckmann and Doebeli, 1999), so it is of considerable biological importance to be able to assess whether a trait is always under stabilizing selection, even though some individual estimates may be compatible with disruptive selection owing to estimation error. Third, we compare different models (with and without stabilizing selection, environmental fluctuations, trends and effects of environmental covariates) using a model selection procedure (Spiegelhalter et al., 2002).

In the following, after introducing the underlying biological model and its statistical formulation, we use simulations of a randomly fluctuating optimum to assess the performance of our approach in detecting key aspects of the model (stabilizing selection, fluctuating selection, autocorrelation of the optimum). We then apply our method to a longitudinal dataset of breeding time in great tits (Reed et al., 2013a; Visser et al., 1998), illustrating how our approach can shed new light on classic examples of selection in response to climate change.

Methods

MODEL

We present our approach in the context of temporally changing selection, and address spatial variation in the Discussion. Our method can be applied to any episode of viability or fertility selection, or to overall selection through lifetime fitness for organisms with discrete non-overlapping generations. For the presentation of the method, we mostly focus on models for fecundity components of fitness, such as number of surviving offspring. Such discrete non-negative variables are best modelled by distributions such as the Poisson, or Poisson mixtures. Within the framework of generalized linear models, the logarithm is a commonly used link function relating the expected value of such response variables to

the linear predictor containing the effects of covariates of interest. If including both linear and quadratic effects of the traits, this leads to a Gaussian model of stabilizing selection.

Writing the model first in terms of parameters of biological interest, the fitness (expected number of offspring) in year t of individuals with phenotype z is given by

$$W_t(z) = W_{\max,t} \exp \left\{ -\frac{(z - \theta_t)^2}{2\omega^2} \right\}, \quad (1)$$

where θ_t is the optimum phenotype at time t , and ω is the width of the fitness peak (smaller ω causes stronger stabilizing selection). We shall assume that θ_t is linearly dependent on one or several observed or latent environmental processes. Letting x_t denote an observed environmental variable, and ϵ_t a latent environmental process,

$$\theta_t = A + Bx_t + \epsilon_t. \quad (2)$$

We assume that ϵ_t is an autoregressive process with autocovariance function $\text{Cov}(\epsilon_t, \epsilon_{t+h}) = \alpha^{|h|}\sigma_\epsilon^2$ and mean $E(\epsilon_t) = 0$. (We also briefly consider a random walk of the optimum, see below).

The primary parameters of fluctuating selection that we need to estimate are thus the width ω the fitness function, and parameters of movements of the optimum θ_t , namely A , B , σ_ϵ^2 and α . These parameters relate to classic theory of evolution under stabilizing selection. For instance, the strength of stabilizing selection can be measured by the reduction in phenotypic (and additive genetic) variance within a generation, which is proportional to the opposite of the curvature of the log-mean fitness landscape. For a Gaussian fitness function, this equals $S = 1/(\omega^2 + \sigma_z^2)$, where σ_z^2 is the phenotypic variance before selection (Phillips and Arnold (1989a); Lande and Arnold (1983); note that S should not be confused with the selection differential, sometimes denoted S or s). It can be shown that $1/S = \omega^2 + \sigma_z^2$ is the squared width of the fitness landscape representing mean fitness as a function of the mean phenotype (Lande, 1976). The quadratic selection gradient estimated by the Lande and Arnold (1983) approach is (in univariate form) $\gamma = S + \beta^2$. Thus, unlike S , γ does not measure the strength of stabilizing selection directly but must be combined with the estimated selection gradient β , as previously pointed out by other authors (Phillips and Arnold, 1989b; Lande and Arnold, 1983; Arnold et al., 2001;

Schluter, 1988). With a Gaussian fitness function, the directional selection gradient as estimated by the Lande and Arnold (1983) approach is $\beta = S(\theta - \bar{z})$, so the response to selection by the mean phenotype is proportional to the strength of stabilizing selection and the deviation of the mean phenotype from the optimum. Predictions for the lag load or evolution of plasticity in a fluctuating environment in turn depend on parameters of fluctuations in θ (e.g., Charlesworth (1993); Lande and Shannon (1996); Chevin (2013); Tufto (2015); Zhang (2012)).

For the statistical estimation of these primary parameters, it is convenient to rewrite equations (1) and (2) as

$$\ln W_t(z) = \mu_t + (\beta_z + \zeta_t)z + \beta_{zz}z^2 + \beta_{xz}x_tz, \quad (3)$$

where the expected number of offspring $W_t(z)$ is treated as the mean parameter of a Poisson distribution, or related distributions such as the negative binomial (a gamma-Poisson mixture). This corresponds to a generalized linear mixed model (hereafter GLMM) with a logarithmic link function. The linear predictor of this GLMM includes (i) a fixed effect of t , treated as a factor (μ_t); (ii) a linear effect of z , with regression slope that includes both a fixed effect β_z and a random effect ζ_t ; (iii) a quadratic effect of z ; and (iv) an interaction between x_t and z . The random effect on the regression slope of z is assumed to have autocovariance function $\text{Cov}(\zeta_t, \zeta_{t+h}) = \alpha^{|h|}\sigma_\zeta^2$. Equation (3) thus describes a form of random regression, where the random effects on the slope are structured according to a first order autoregressive process (AR1). Equation (3) also is a latent Gaussian model, a class of hierarchical models where an unobserved latent Gaussian process, with distribution described by some hyperparameters, gives rise to observations that are independent conditional on the latent process. Here, the latent process is fluctuating selection, that is, the changing trait-fitness relationship described by the slopes in eq. (3) and defined at the population level, whereas the observations are individual phenotypes and fitnesses at each time point.

The primary parameters of fluctuating selection can be recovered from this GLMM

through the transformations

$$\begin{aligned}\omega &= \sqrt{-\frac{1}{2\beta_{zz}}}, & \theta_t &= -\frac{\beta_z + \beta_{xz}x_t + \zeta_t}{2\beta_{zz}}, & B &= -\frac{\beta_{xz}}{2\beta_{zz}}, \\ \sigma_\epsilon^2 &= \frac{\sigma_\zeta^2}{4\beta_{zz}^2}, & W_{\max,t} &= \exp\left\{\mu_t - \frac{(\beta_z + \beta_{xz}x_t + \zeta_t)^2}{4\beta_{zz}}\right\}.\end{aligned}\quad (4)$$

In our analyses, we also fit a GLMM similar to (3) but with no quadratic term ($\beta_{zz} = 0$), in order to compare models with and without stabilizing selection. This GLMM corresponds to an exponential fitness function, for example, $W_t(z) = W_{\max,t} \exp\{\mu_t + (\beta_z + \zeta_t)z\}$, for which the selection gradient in year t is $\beta_z + \zeta_t$ regardless of the phenotype distribution. Such a fitness function is thus a good null model for fluctuating selection, since all variation in selection gradients must come from environmental changes affecting β_t .

When the selection episode under investigation is based on viability rather than fecundity selection, the number of surviving individuals is best modelled as a binomially distributed variable, or a betabinomial to account for overdispersion. The survival probability $w(z)$ can then be related to covariates of interest through a logit link function $\ln(p/(1-p))$ (logistic regression), as in [Janzen and Stern \(1998\)](#), or a complementary log log link ($\ln(-\ln(1-p))$), which arises naturally in survival analysis with a constant hazard rate (instantaneous mortality risk). Both link functions produce symmetric stabilizing selection, but the sigmoid shape of these functions necessarily implies a flattening of the fitness function and weaker selection for high survival rates close to one, because high survival implies less opportunity for selection to occur. This also makes the overall fitness function non-Gaussian, although both fitness functions are well approximated by a Gaussian for low survival.

INFERENCE AND MODEL SELECTION

All inferences were made using INLA ([Rue et al., 2009](#)), an R package for Bayesian statistics, which uses integrated nested Laplace approximation for the rapid computation of posterior densities for latent Gaussian hierarchical models. The rationale behind our choice of priors is discussed in details in Appendix [S1](#). In brief, we relied on non- or weakly informative priors for most parameters, except for ω when the latter was used in

computations that assume stabilizing selection, in which case we sometimes used a prior that assumes that $\beta_{zz} < 0$ in the case study.

Our approach relies on comparison between models with different degrees of complexity and number of parameters. Regardless of the actual underlying process, the quality and quantity of available data should determine whether the best model is one that includes temporal changes in the optimum, or a simpler one that considers a constant fitness function (with or without an optimum). In order to compare the fit obtained with different models, we use the deviance information criterion (hereafter DIC), an analog to Akaike’s information criterion (AIC) for hierarchical Bayesian models (Spiegelhalter et al., 2002). Both the AIC and the DIC are model selection criteria based on information theory (Burnham and Anderson, 2002), which allow ranking the predictive power of different models using their likelihoods, penalizing models with more parameters. Among the models considered, the model with the lowest DIC is chosen, and models within DIC 2 to 3 points larger than the best model are considered to have nearly as strong support from the data as the best model (Spiegelhalter et al., 2002; Burnham and Anderson, 2002). Using other newly developed criteria, the Watanabe-Akaike information criterion (Gelman et al., 2014), as well as a cross-validation score based on conditional predictive ordinate-values (Gneiting and Raftery, 2007; Held et al., 2010) yielded very similar model rankings (Figure S3.1).

SIMULATIONS

The ability of the method to identify the best model was tested using simulations of fluctuating selection. We modeled samples from a population undergoing stabilizing selection with autocorrelated fluctuations in the optimum, as in equations (1) and (2). The sample size was drawn each year from a Poisson distribution with mean $n = 25$ or 50 individuals. We considered 40 time points, which is a large but still realistic number that compares to existing long-term studies of natural populations (e.g. Reed et al., 2013a; Vedder et al., 2013; Grant and Grant, 2002). We drew individual phenotypes from the same normal distribution each year, that is, we neglected responses to selection and ge-

netic drift for simplicity. Indeed, our approach estimates selection based on trait-fitness relationships in each time, and does not use information on responses to selection. However, responses to selection may buffer or amplify deviations of the mean phenotype from the optimum (Lande and Shannon, 1996; Chevin, 2013; Tufto, 2015), in effect changing the magnitude of environmental fluctuations, which may affect the estimation procedure to some extent (see effect of σ_θ in Figure 1b). The phenotypic variance before selection σ_z^2 was set to 1 without loss of generality, which is equivalent to focusing on standardized trait values scaled to their phenotypic standard deviation. For each individual, its fitness was computed from its phenotype using (1), and its actual number of offspring was then drawn from a Poisson distribution with mean $W_t(z)$. We varied the width of the fitness function ω such that the strength of stabilizing selection $S = 1/(\omega^2 + \sigma_z^2)$ (e.g. Lande, 1976) was 0.025, 0.05, 0.1, or 0.2. For each width of the fitness function, the standard deviation (hereafter SD) of fluctuations in the optimum was varied from 0 (constant optimum) to 0.5, 1, 2, and 4. This combination of widths of the fitness function and magnitudes of fluctuations in the optimum causes $S\sigma_\theta$, the SD of directional selection gradients without response to selection and neglecting genetic drift, to vary from 0 to 0.8. For each combination of these parameters, we varied the autocorrelation of the optimum, $\alpha = 0$ (no autocorrelation), 0.1, 0.25, 0.5, 0.75, 0.9. We also allowed temporal changes in fitness beyond those caused by fluctuating selection, by drawing the maximum fitness $W_{\max,t}$ from a log normal distribution with mean 5 and SD 0.2 on the log scale. The mean number of offspring across years was thus 5.1 for hypothetical individuals with the optimum phenotype. We also ran simulations with an exponential fitness function $W_t(z) = W_{\max,t} \exp\{\beta_t z\}$, which corresponds to pure directional selection (no stabilizing selection) at each time, with gradient β_t . In those simulations, we drew β_t for each time point from a normal distribution with mean 0 and SD ranging from 0 to 0.8 (by increments of 0.2), thus covering the same range of variation in directional selection gradients as in simulations with a moving optimum. We used the same values for the autocorrelation of β_t as those for θ_t .

We then ran the estimation procedure under a choice of models. Because the number of

parameters of the complete model in equation (4) is relatively large, we did not consider all possible models but instead focused on a subset of biologically meaningful models to test particular questions, as recommended for model selection (Burnham and Anderson, 2002). Specifically, we fitted (i) the AR1 model in equation (4), with no covariate x ; (ii) a similar model but with uncorrelated fluctuations (white noise), i.e. random Poisson regression with independent random effects on the slope; (iii) a model with autoregressive slope but no quadratic term (no optimum); (iv) a model with a constant optimum (quadratic term and no random effect on the slope), implying no fluctuating selection. For each parameter set, we ran 200 simulations, and we computed the proportion of simulations where a given feature of the models was selected based on DIC, as a way to quantify the support for this feature. For instance, the proportion of simulations where any model with a quadratic term has the lowest DIC by at least 2 points measured support for stabilizing selection, and the proportion of simulations where any of the models with a random effect and a quadratic term was chosen based on DIC measured support for a fluctuating optimum.

CASE STUDY ON GREAT TITS

We then applied our method to a dataset of fecundity selection on breeding date in the Hoge Veluwe population of great tits (*Parus major*), in the Netherlands. We used years from 1973 to 2013, and removed all clutches that had been manipulated experimentally, following Reed et al. (2013b). The remaining dataset included 3782 clutches. The trait z we investigated was breeding date (date of first egg laying), and we measured the fecundity component of fitness W as the number of individuals surviving to become fledglings. We could have used number of recruits surviving to the next breeding season, as in Reed et al. (2013a), but this has much smaller mean and hence larger coefficient of variation due to demographic stochasticity, causing more uncertainty in estimates of all parameters and reducing statistical power. A somewhat high probability of clutch failure (around 32%, see below) caused the distribution of clutch sizes to be manifestly zero-inflated (Supplementary Figure S3.2), so we modeled a zero-inflated Poisson instead of a Poisson distribution as above, unless otherwise stated. To account for possible non-independence

in the data caused by some females reproducing repeatedly in different years, we also investigated models with female identity included as a random effect in the analysis. To account for few replicates for a large proportion of females we used a correction described in [Ferkingstad and Rue \(2015\)](#). Our dataset also included environmental variables that we used as covariates to explain movements of the optimum: the mean temperature over a critical time window during which it influences the date of the peak in food biomass and, when available, the date of the food peak itself ([Visser et al., 2006](#)).

Results

EVALUATING THE METHOD WITH SIMULATIONS

We first tested our approach on simulations of an optimum phenotype undergoing an autoregressive process. Our first aim was to assess whether our method is efficient in detecting stabilizing selection (assuming constant ω), regardless of patterns of fluctuations. Figure (1a) shows that the proportion of simulations where the best model (by at least 2 DIC points) includes a quadratic term increases rapidly with the actual simulated strength of stabilizing selection, and approaches 1 for S as small as 0.05. This corresponds to weak stabilizing selection ([Lande, 1975](#)), which would typically be non-significant if estimated from a single time point with moderate sample size as here (see Figure 8 in [Kingsolver et al. \(2001\)](#), which reports estimates of the standardized quadratic gradient $\gamma = S + \beta^2 \approx S$ under weak directional selection). Here, stabilizing selection is detected more efficiently by pooling data from several time points.

Figure (1b) shows that the proportion of simulations where the best model includes fluctuations in the optimum phenotype increases with the strength of stabilizing selection (darker lines indicate larger $S = 1/(\omega^2 + \sigma_z^2)$). In particular, simulations with fluctuating selection but no optimum (lighter lines) essentially do not lend support to any model with a moving optimum. For a given S , the probability to detect a fluctuating optimum increases with the magnitude of actual fluctuations in this optimum, except in the case of large fluctuations, under strong selection (rightmost point in darkest lines). In this case,

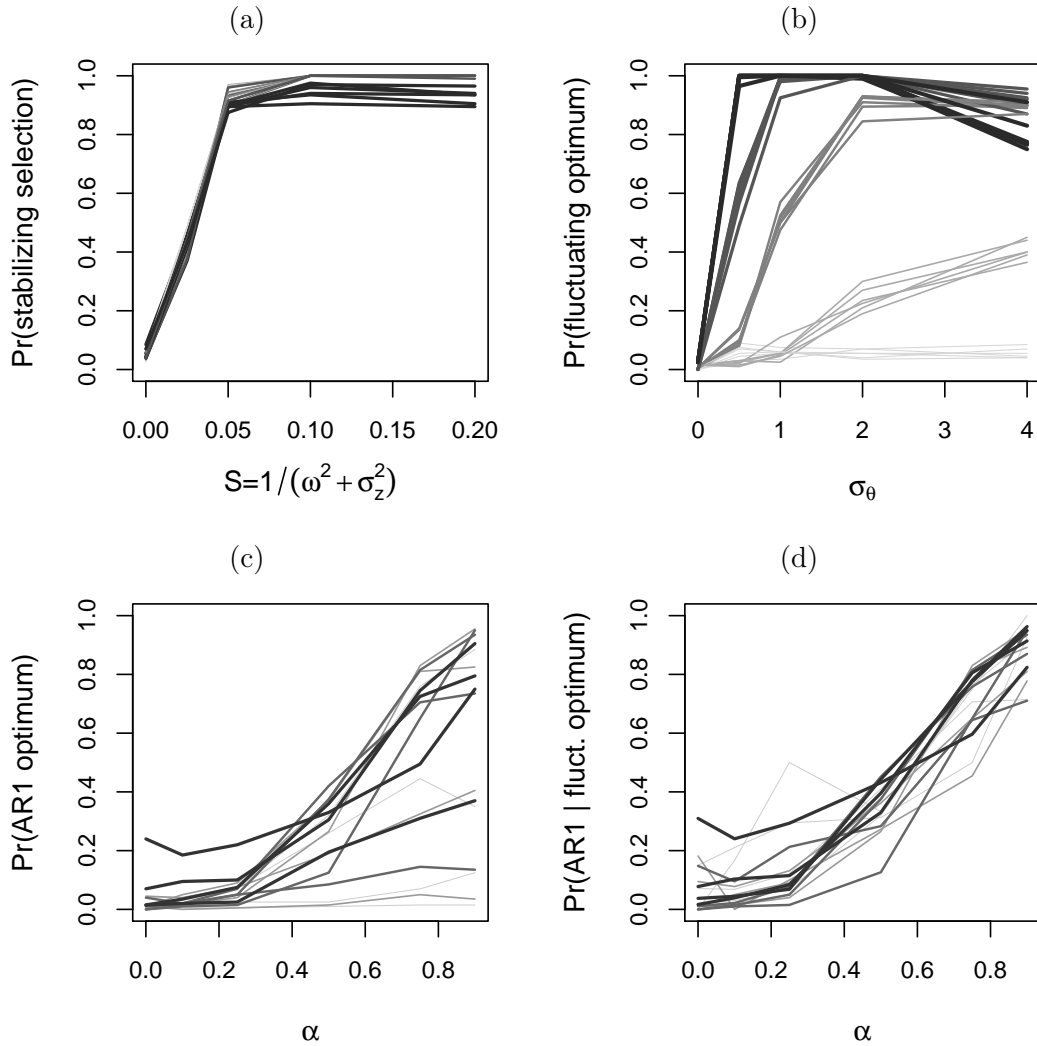


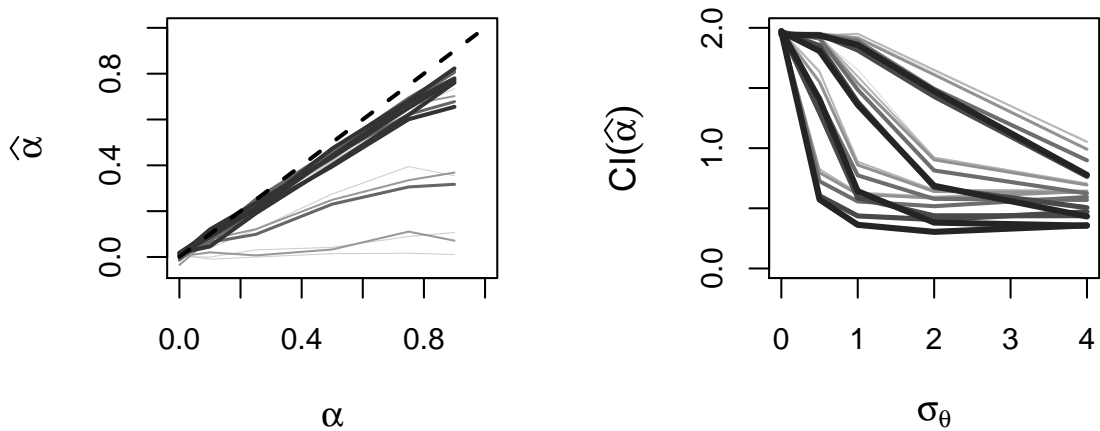
Figure 1: **Test of the approach with simulations.** The proportion of simulations of fluctuating selection (out of 200 repeats) where the model that best fits the data by at least 2 DIC points includes a specified feature of selection (below, where this feature is ‘detected’), is shown against the actual parameters used in simulations. (a) Detection of stabilizing selection (quadratic term in (3)), against the strength of stabilizing selection $S = 1/(\omega^2 + \sigma_z^2)$. Values for $S = 0$ correspond to simulations with $W_t(z) = W_{\max,t} \exp\{\beta_t z\}$, which include no stabilizing selection. (b) Detection of any form of random fluctuation in an optimum phenotype, against the actual standard deviation σ_θ of simulated fluctuations in the optimum. (c) Detection of an autoregressive (AR1) optimum, against the actual autocorrelation α of simulated fluctuations in the optimum. (d) Detection of an AR1 optimum, conditional on detecting a fluctuating optimum. The gray scale and line width represents the magnitude of fluctuations in (a), (c) and (d), with darker and thicker lines indicating larger σ_θ , and the strength of stabilizing selection in (b). The lighter and thinner lines in (b) correspond to the exponential fitness function (with no optimum), while only simulations with a fluctuating optimum were used in (c) and (d). Simulations were run for 40 time points, with the mean sample size per time point being $n = 50$, the phenotypic variance $\sigma_z^2 = 1$, the width of the fitness function such that $S = 0.025, 0.05, 0.1, 0.2$. Fluctuations in the optimum are determined by their standard deviation $\sigma_\theta = 0$ (constant optimum), 0.5, 1, 2, 4, and autocorrelation $\alpha = 0$ (white noise), 0.1, 0.25, 0.5, 0.75, 0.9.

fluctuating selection is strong enough to substantially reduce the expected number of offspring of most individuals in most generations, reducing the power to detect variation in selection as compared to more moderate fluctuations. The probability to detect fluctuating selection when the optimum is actually constant (values for $\sigma_\theta = 0$ in fig (1b)), which has some connection to type I error, was always low in our simulations.

The probability to detect an autocorrelated optimum (AR1 process) increases with the actual autocorrelation of fluctuations, but also with their magnitude as indicated by the gray scale (Figure (1c)). AR1 is detected in at most 20% of simulations for $\alpha < 0.4$, and even strongly autocorrelated fluctuations are not detected as AR1 if their magnitude is small (lighter lines in (1c), for which $\sigma_\theta = 0.25, 0.5$). However, conditional on detecting a fluctuating optimum (either white noise or autoregressive), the proportion of simulations that do detect autocorrelation becomes substantial and increases with increasing α (Figure 1d).

The autocorrelation of the optimum is estimated more accurately under larger fluctuations (larger σ_θ): the posterior mean of α is closer to its actual value for dark as compared to light lines in Figure 2a, and the 95% credible interval becomes narrower towards the right in Figure 2b. The credible intervals on α are very sensitive to the strength of stabilizing selection, being reduced by a factor of 3 as S changes from small to large (light gray to dark lines in Figure 2b). But in all cases the credible intervals in Figure 2b remain quite large, showing that our simulations with 40 time points do not provide sufficient information to estimate the autocorrelation of the optimum with high accuracy, even in cases where an autocorrelated optimum may be detected in Figure (1c). Similarly, the posterior mean α is slightly biased towards 0, the mean of the prior distribution, even under large fluctuations (Figure 2a), as expected generally when limited sample size does not allow the likelihood to completely dominate the prior.

Our hierarchical Bayesian method estimates the joint posterior distribution for all parameters, from which the marginal posterior distribution for each time point can be recovered. The primary parameters of our biological model depend non-linearly on parameters in the GLMM (see eq. (4)). However, their posterior marginal distribution can



(a) Estimated α

(b) Credible interval for α

Figure 2: **Inferred autocorrelation of optimum in simulations.** (a) The posterior mean for the autocorrelation of movements of the optimum phenotype are plotted against the actual autocorrelation α used in simulations. Darker and thicker lines indicate larger fluctuations in the optimum (as quantified by σ_θ), and the dashed line plots $y = x$. (b) The width of the 95% credible interval for α is plotted against the magnitude of fluctuations in the optimum, quantified by σ_θ . Darker and thicker lines indicate stronger stabilizing selection (larger S). Parameters as in Figure 1a.

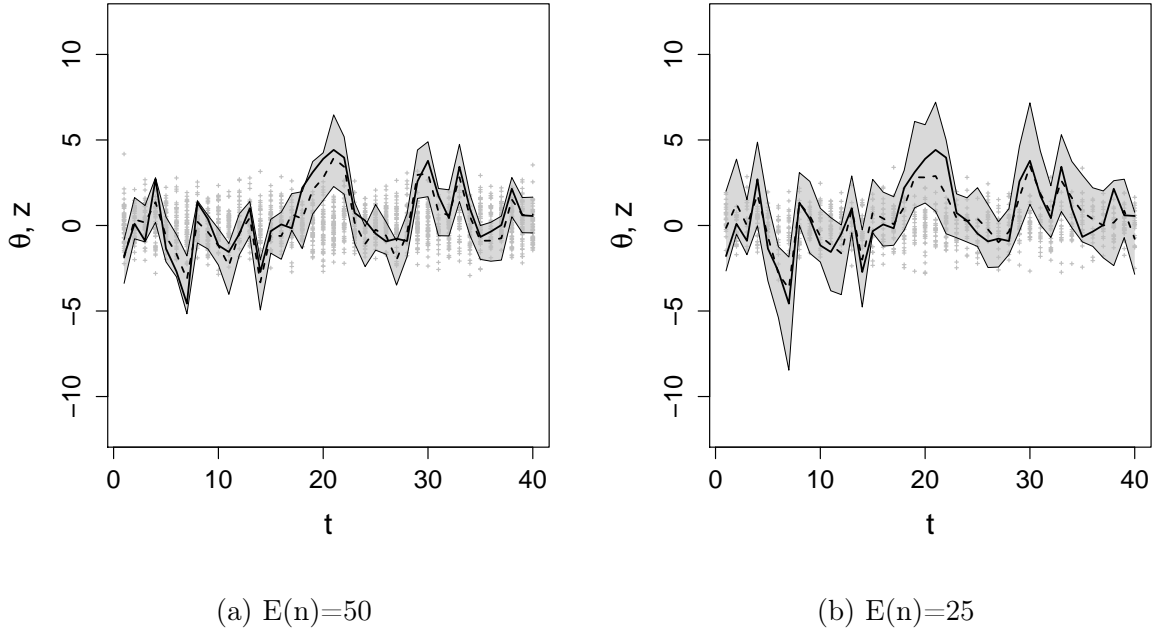


Figure 3: **Estimates of fluctuating optimum in an individual simulated sample.** The simulated movements of an optimum phenotype (thick full line) are represented along time in a particular simulation repeat. The posterior mean for the estimated optimum is also represented in dashed, together with the 0.95 credible interval (shading). The dark gray dots are the individual phenotypes. The same simulated pattern of phenotypic selection is estimated with mean sample size per time point $n = 50$ (left) or 25 (right). The phenotypic variance is $\sigma_z^2 = 1$, the width of the fitness function is $\omega = 3$, the optimum phenotype fluctuates with $\sigma_\theta = 2$ and $\alpha = 0.5$, and all other parameters are as in Figure 1a.

be found by drawing samples from the joint posterior distribution of parameters of the GLMM (as allowed by packages such as INLA), and applying the transformation (4) to each sample. An example of estimated fluctuations of an optimum phenotype is illustrated in Figure 3. When the mean sample size is $n = 50$, the movements of the optimum are inferred with good accuracy, including outside of the current phenotypic range (individual phenotypes appear as gray dots), and the 95% credible interval includes the true optimum at all time points. Reducing the mean sample size to 25 widens the credible interval and increases the error of the posterior mean, making it closer to the long-term mean, but the major tendencies of the optimum (such as the sign of the deviation from the mean optimum) are still well captured in this example.

CASE STUDY: PHENOLOGY OF GREAT TITS

To illustrate the usefulness of our approach, we next apply it to a well-known example of climate-change induced phenotypic selection: breeding time in the Great tit population of Hoge Veluwe in the Netherlands (Reed et al., 2013a; Visser et al., 1998; Reed et al., 2013b). In this species, there is strong evidence that the date of egg laying is under selection for an optimum set by the peak in the biomass of caterpillars that are the main food resource for hatchlings (Visser et al., 2006). But movements of the optimum phenotype have not been estimated directly from the phenotypes and fitnesses of all individuals in this population.

Table 1 shows a number of candidate models for the Hoge Veluwe great tit data. We first fitted models without any environmental covariates. Given this model form, there was weak evidence for a nonzero autocorrelation between the optima in different years t (DIC reduced by 3.29 for model 2 vs. 1, with ζ_t AR1 or iid, respectively). We investigated whether the data showed evidence for non-stationarity, by fitting an alternative model where the optimum follows a random walk (discrete time equivalent of Brownian motion as modeled by Estes and Arnold (2007) and Hansen et al. (2008)) instead of AR1, but this increased the DIC by 1.4 (model 3 versus model 2). Fitting an alternative model with a linear trend in the optimum, equivalent to an environmental covariate $x_t = t$, did not reduce the DIC either (models 4 and 5, with ζ_t iid or AR1, respectively).

Selection on laying date in great tits (and other passerines) is known to depend on the timing of the peak in the biomass of caterpillars which they use as food (Visser et al., 2006). This in turn partly depends on the phenology of trees, which is largely determined by spring temperature. Having estimates of the mean temperature over a critical time window known to influence the food peak (Visser et al., 2006), we changed the model to include this as a covariate x_t instead of a linear trend in the optimum, but still allowing for random fluctuations beyond those caused by spring temperature. This improved the model, causing DIC to decrease by 4.21 for models with AR1 residual variance (model 7 vs. 2). With this environmental covariate included, there was still evidence for nonzero autocorrelation in the remaining variation in the optimum (ϵ_t AR1 instead of iid) as shown by the decrease in DIC by 4.9 when comparing model 7 to model 6.

Table 1: **Model selection for temporal variation in selection on laying date in great tits.** Linear predictors are shown for different candidate models of selection on breeding time, together with the model assumed for the random effect (ζ_t being either independent and identically normally distributed, iid, or following an autoregressive order 1 process, ar1). Also shown are the associated differences in DIC (Δ DIC) relative to the best model (model 7). Further details are given in the main text. The number of fledglings is obtained from the linear predictor as a zero-inflated Poisson variable, except in models 13 and 14 (Poisson and zero-inflated negative Binomial, respectively). Mother identity was included as a random effect in model 15, treating some cases where the mother was unknown as different mothers.

Model	Linear predictor: $\mu_t + \beta_z z +$	ζ_t	Δ DIC
1) No covar.	$\beta_{zz} z^2 + \zeta_t z$	iid	7.5
2) No covar.	$\beta_{zz} z^2 + \zeta_t z$	ar1	4.21
3) Random walk	$\beta_{zz} z^2 + \zeta_t z$	rw1	5.61
4) Linear trend	$\beta_{zz} z^2 + \beta_{zt} z t + \zeta_t z$	iid	7.61
5) Linear trend	$\beta_{zz} z^2 + \beta_{zt} z t + \zeta_t z$	ar1	4.43
6) Temper.	$\beta_{zz} z^2 + \beta_{zx} z x_t + \zeta_t z$	iid	4.9
7) Temper.	$\beta_{zz} z^2 + \beta_{zx} z x_t + \zeta_t z$	ar1	0
8) No random slope	$\beta_{zz} z^2 + \beta_{zx} z x_t$		32.4
9) Temper. + trend	$\beta_{zz} z^2 + \beta_{zx} z x_t + \beta_{zt} z t + \zeta_t z$	ar1	0.24
10) No stab. sel.	$\beta_{zx} z x_t + \zeta_t z$	ar1	25.18
11) Free lin. slopes	$\beta_{zz} z^2 + \beta_{zt} z$		14.55
12) Free lin. and quadr.	$\beta_{zz} z^2 + \beta_{zt} z + \beta_{zzt} z^2$		14.19
13) No. zero.infl.	$\beta_{zz} z^2 + \beta_{zx} z x_t + \zeta_t z$	ar1	6129.08
14) Zero. infl. neg. bin.	$\beta_{zz} z^2 + \beta_{zx} z x_t + \zeta_t z$	ar1	90.96
15) Fem. id. rand. eff.	$\beta_{zz} z^2 + \beta_{zx} z x_t + \zeta_t z$	ar1	8.2

We then compared the model with the spring temperature and autocorrelation included (model 7) to a number of alternatives, including a model without random fluctuations in the optimum beyond those generated by spring temperature (model 8). Removing the random fluctuations led to an increase in DIC of 32.4, strongly suggesting that additional environmental variables other than spring temperature itself influence

the optimum. A model including both a linear trend in the optimum and effects of the spring temperature was also considered (model 9) but this model did not improve DIC. Comparing model 7 with a model without the quadratic term $\beta_{zz}z^2$, that is, an exponential fitness function with random fluctuations in the slope of log fitness on the trait (model 10), strongly favoured our selected model based on stabilizing Gaussian selection (difference in DIC equal to 25.18), where directional selection comes from mismatches between the mean phenotype and the optimum.

We also considered models with the regression slopes β_{zt} estimated as free parameters in different years (model 11), rather than being modelled as random effects. This model was further extended to also include different quadratic regression coefficient β_{zzt} in different years (model 12). These more flexible model alternatives, providing a form of goodness-of-fit test of our model, did not lead to any improvement in DIC. Thus we conclude that there is no evidence in the data for any deviations in the local slope and curvature of the fitness function beyond what is predicted by our Gaussian model with a fluctuating optimum (1).

The last three models were used as tests for basic features of our statistical model. Our choice of residual distribution of fitness was supported by the strong increase in DIC when replacing the zero-inflated Poisson by a regular Poisson distribution (model 13), or by a zero-inflated negative binomial (model 14). This last comparison suggests that there is little overdispersion in the number of fledglings produced by a given female in a given year, beyond that caused by the zero-inflation. DIC also increased (by 8.2) if including a random effect associated to female identity, indicating that there was no significant variation in annual number of fledglings between females.

Parameter estimates under our selected model (model 7) and the best model without any environmental covariates (model 2) are listed in Tables 2 and 3, respectively. Both models estimate 32% zero inflation (p_0). The inclusion of temperature as a covariate greatly improved the precision of estimates for several of our model parameters (smaller S.E. and CI in Table 2 than in Table 3). This improvement agrees well with our intuition about how our method extracts information about the different parameters from the

Table 2: Estimate of the width ω of the Gaussian fitness function, the intercept A and slope B in the regression of the optimum θ_t on temperature x_t and the standard deviation σ_ϵ^2 and autocorrelation α of additional fluctuations in the optimum, for the selected model of selection on onset of breeding through the number of fledglings (model 7 in Table 1). The parameter estimates were computed by transforming samples from the posterior according to eqs. (4). Also reported are the zero-inflation parameter p_0 , and the standard deviation σ_ν of the random effect on maternal identity.

Parameter	Posterior mean \pm S.E.	95% credible interval
ω (days)	20.55 ± 1.7	(17.85, 24.42)
σ_ϵ (days)	6.75 ± 1.66	(4.4, 10.53)
Autocorrelation α	0.3029 ± 0.2419	(-0.2176, 0.7113)
Intercept A (April day)	19.43 ± 1.95	(14.97, 22.89)
Slope B (days/ $^\circ$ C)	-5.01 ± 1.09	(-7.38, -2.93)
p_0	0.32 ± 0.01	(0.31, 0.33)

Table 3: Same as Table 2, but for the model without environmental covariates (model 2 in Table 1).

Parameter	Posterior mean \pm S.E.	95% credible interval
ω (days)	24.11 ± 2.73	(20.08, 30.78)
σ_ϵ (days)	11.3 ± 2.56	(7.5, 17.94)
Autocorrelation α	0.2472 ± 0.2075	(-0.1745, 0.626)
A (April day)	17.66 ± 2.95	(11.3, 23.06)
p_0	0.32 ± 0.01	(0.31, 0.33)

data. For instance regarding ω , with some - albeit incomplete - information about current location of the optimum provided by the environmental covariate at different time points, model 7 is better able to infer the strength of stabilizing selection from the local steepness of the fitness function. In contrast, for model 2 not utilizing this information, the strength of stabilizing selection can to a greater extent only be inferred from the local curvature of the fitness function, making the uncertainty in ω much greater.

The estimated width of the fitness function in the best model was $\omega = 20.55$ days. With the average within-year phenotypic standard deviation being $\sigma_z = 5.35$ days, this yields for the standardized strength of stabilizing selection on the unitless trait measured in the scale of phenotypic standard deviation, $S' = \sigma_z^2/(\omega^2 + \sigma_z^2) = 0.063$, which can be compared directly to the S used in simulations (see Figure (1a)). The estimated environmental sensitivity of selection, as quantified by the slope B of the optimum on the environmental covariate (here temperature), is such that the optimum laying date advances by 5.01 days/ $^{\circ}\text{C}$. In model 2 that does not include temperature as a covariate, A is simply the mean optimal laying date across years, which is estimated to be April 18 (Table 3) whereas in model 7 A is the optimal laying date (April 19) at the mean value of spring temperature (since most of our covariates were mean-centered). Finally, note that the standard deviation of the optimum is estimated to be larger in model 2 than in model 7 (11.3 versus 6.75 days, respectively), as expected because in the latter model part of the fluctuations are captured by the relationship with temperature, and σ_{ϵ} only describes variation that is left after accounting for this.

Having established an effect of the spring temperature x_t on the optimum, we analysed the pattern of fluctuations in this variable. Direct observations of temperature yields more power to detect autocorrelation, but after removal of the long term trend in x_t by ordinary linear regression, there was no evidence for autocorrelation in the residuals of x_t . The standard deviation σ_{x_t} around the long term trend was estimated to 1.07°C with a 95%-confidence interval of (0.87, 1.37). This gives for the standard deviation of fluctuations in the optimum, arising from the combined effect of variation in spring temperature and the latent random variable ϵ_t , $\sigma_{\theta} = \sqrt{B^2\sigma_{x_t}^2 + \sigma_{\epsilon}^2} = 8.62$ days. A proportion $B^2\sigma_{x_t}^2/\sigma_{\theta}^2 = 38\%$ of the total variance in θ_t is thus attributable to fluctuations in temperature around the long term trend.

The estimated patterns of fluctuations in the optimum are represented in Figure 4. The right panel shows how the fitness function scales in comparison to patterns of phenotypic variation in the trait, represented by a Gaussian curve (both curves are standardized to the same height). The average phenotype distribution (represented by the gray curve)

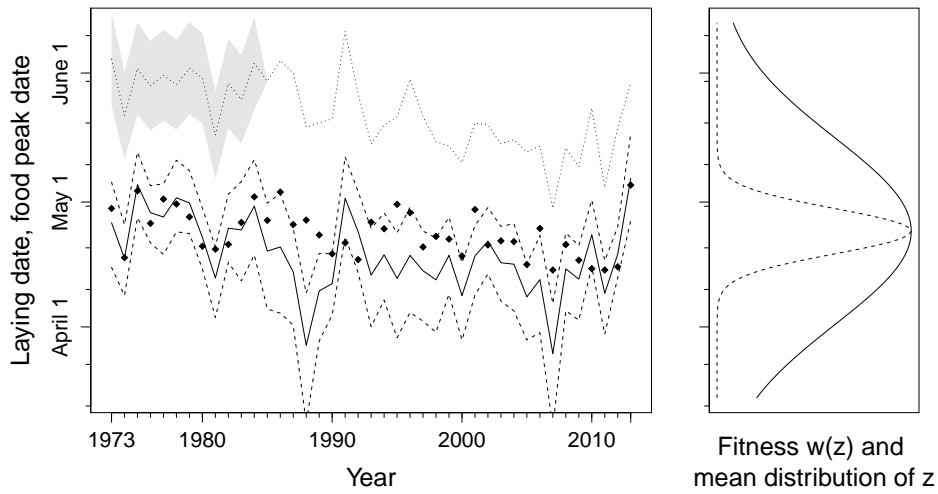


Figure 4: **Estimated movements of the optimum laying date.** The time series of estimates for the optimum phenotype θ_t (solid line), based on the selected model for number of fledglings, is shown on the left panel, together with its 95% credible interval (dashed lines). Also shown is the mean laying date in each year (dots), and the timing of the peak in caterpillar biomass (dotted line). For years before 1985, the food peak date was not available, so an estimate based on linear regression on spring temperature was used, with a 95%-prediction interval represented by the shaded zone. The right panel shows the estimated Gaussian fitness function in solid curve, and a normal density representing the mean within-year phenotypic distribution (variance equal to mean empirical variance in onset of breeding across all years) in dashed curve, both centered on the mean laying date across years.

is narrower than the fitness function, consistent with the estimated weak stabilizing selection. The time series of estimates of θ_t are shown in the left panel (mean of the posterior marginal for each date in continuous line, 95% credible interval in dashed lines), using the same scale for the Y axis as in the right panel. This gives a sense of how much the optimum fluctuates relative to the widths of both the fitness function and the phenotype distribution. Interestingly, the mean laying date (represented as dots in Figure 4) is later than the optimum in most years (by an average 5.1 days, based on posterior mean estimates of θ_t each year), resulting in a biased mismatch with optimum and directional selection for earlier breeding. Figure 4 also shows the date of the peak in the biomass of caterpillar food (dotted line). The latter is either directly available (after 1985), or predicted from the spring temperature (before 1985), using a linear model based on years where measurements exist (Visser et al., 2006). It is striking that the optimum phenotype closely tracks major movements of the food peak, but with some advance. To investigate this further, we re-analysed the data using the food peak date as a covariate in addition to spring temperature, for years where this date was available (Appendix S2). Using the food peak date as a covariate and assuming a constant lag between the optimum and the food peak yielded a better model than using any other covariate in isolation (temperature, or time for linear trend, models 8 to 11 in Table S2.1). However, the best model included all three covariates (model 1 in Table S2.1). In models that assume that the optimum laying date occurs a fixed number of days before the food peak at that no other covariate affect the optimum (model 10 in Table S2.1), the mean lag is about 33 days (Table S2.2).

Discussion

IMPLICATIONS OF THE APPROACH AND RESULTS

The variance of phenotypic selection across time or space cannot be estimated from individual estimates of selection gradients (or differentials), because this conflates the true variance with sampling variance (as emphasized by Morrissey and Hadfield, 2012). In-

stead, it can be estimated through a random effect (across years or locations) on the regression slope of relative fitness on individual phenotype. Allowing a flexible covariance structure for this random effect makes it possible to model different forms of autocorrelation in phenotypic selection. Here we assumed a first-order autoregressive process (AR1) for phenotypic selection (and we also fitted a random walk), but our method can easily accommodate other structures, such as ARp (p-order autoregressive process), or continuous-time processes such as white noise, Ornstein-Uhlenbeck, or Brownian motion (as in e.g. [Arnold et al. \(2001\)](#); [Hansen et al. \(2008\)](#)).

The linear regression method of [Lande and Arnold \(1983\)](#) could be directly extended to the measurement of fluctuating selection, by including a random effect on the linear (directional) gradient β , thus measuring the variance and autocorrelation of directional selection. Linear regression with a random effect on the slope in the form of a stochastic process is sometimes described as dynamic linear regression ([Petris et al., 2009](#)), a class of state-space models ([Shumway and Stoffer, 2010](#)). However, our approach based on GLMM with a log link function proves more useful for several reasons. First, we directly estimate a single parameter ω that determines the strength of stabilizing selection, while the quadratic gradient γ does not. Indeed, movements of an optimum with constant width would produce changes in γ , while they would not affect the strength of stabilizing selection (see Methods, and [Phillips and Arnold \(1989b\)](#); [Lande and Arnold \(1983\)](#); [Schluter \(1988\)](#)). Hence, temporal variation in γ , as reported in some empirical studies (see e.g. Figure 2B in [Charmantier et al., 2008](#)), can be perfectly consistent with constant stabilizing selection (constant ω in our model), but changing deviation of the mean phenotype from the optimum. Second, the use of linear regression for analysis of phenotypic selection has been criticized for its reliance on the non-realistic assumption of Gaussian distribution of fitness residuals, which leads to wrong confidence intervals and hypothesis testing using F- or t-tests, despite providing unbiased estimates of linear and quadratic regression coefficients ([Mitchell-Olds and Shaw, 1987](#); [Janzen and Stern, 1998](#); [Shaw and Geyer, 2010](#); [Morrissey and Sakrejda, 2013](#)). This problem may be circumvented by using non-parametric resampling methods such as bootstrapping or jack-knifing, as first sug-

gested by Lande and Arnold (1983). Here, we chose to use the more realistic statistical framework of generalized linear mixed models (GLMM), which in general makes more efficient use of the data than linear regression.

Instead of inferring the selection gradients *per se*, we use a model of stabilizing selection with a moving optimum, thus providing estimates of a number of parameters appearing in theoretical models of fluctuating selection (Bull, 1987; Charlesworth, 1993; Lande and Shannon, 1996; Bürger and Gimelfarb, 2002; Lande, 2009; Chevin, 2013; Tufto, 2015). This should make results of empirical studies more readily interpretable in terms of these theoretical models, while the variance and autocorrelation of selection gradients themselves generally are not sufficient to infer patterns of fluctuations of an optimum phenotype (Chevin and Haller, 2014). In our approach, the assumption of stabilizing selection can be removed simply by not including the quadratic term in equation (3), which also allows comparing models with and without stabilizing selection (through likelihood-ratio tests or using information criteria, as we did here). Also in line with most theory, we modeled stabilizing selection as a Gaussian fitness function, which approximates well any fitness function with an optimum when sufficiently close to the optimum (Lande, 1976) as in Figure 4. In general, however, the Gaussian fitness function should not be seen as a realistic representation of the actual fitness landscape, which may be better estimated through non-parametric approaches such as cubic splines (Schluter, 1988; Morrissey and Sakrejda, 2013). Rather, it is an idealized fitness function that allows quantitative comparison of patterns of fluctuating selection across studies, as long as stabilizing selection with a moving optimum is biologically meaningful. A perhaps useful analogy is the Wright-Fisher population, which is an idealization that allows comparing the intensity of drift across populations with different complex demographies, through the summary parameter N_e .

Our simulations resulted in rather imprecise estimates of the autocorrelation in the optimum phenotype, with large confidence intervals (Figure 2b)). This is because we restricted our attention to the modest sample sizes and number of time points that are typical of studies on organisms with intermediate to large generation times (mostly

birds and large mammals) in (semi-)natural conditions. The precision for these estimates cannot be increased by only increasing the sample size within time points, and more time points are needed instead. For instance, even with perfect knowledge of the optimum at each time point, the estimate of the temporal variance in the optimum (σ_θ^2 , or σ_ϵ^2 in the case without an environmental covariate) has coefficient of variation $\sqrt{(2/(n-1))}$, such that $n > 2/u^2 - 1$ time points are required in order for the coefficient of variation to be below a given threshold u . This problem is further amplified by the uncertainty in the estimate of each optimum. More precise estimation of patterns of fluctuating selection may be obtained for species with shorter generations, such as those studied in experimental evolution, allowing for measurements over a larger number of time points.

The results from our analysis of the Hoge Veluwe great tit data were partly in agreement with previous findings, but also revealed new and intriguing patterns. We showed that the optimum breeding time responds to spring temperature, which could be anticipated since the latter is a good predictor of the food peak date. However, in years where the food peak date itself was measured, it was not a better predictor of the optimum than spring temperature. Furthermore, we were able to show that the optimum laying date undergoes temporally autocorrelated random fluctuations, beyond those caused by spring temperature. These fluctuations may be caused for instance by other climatic variables than temperatures, other spring food sources that are less dependent on the timing of bud burst in trees, or winter conditions. A possible candidate driver of these fluctuations is the number of sunspots, which has been shown to influence laying date independently of mean spring temperature (Visser and Sanz, 2009), and may thus also affect selection on this trait. Another plausible candidate is the height of the food peak, which determines the total amount of food available in each year, and which strongly affects the number of fledglings produced. It has also been noted that food peak height correlates with the number of sunspots (Visser and Sanz, 2009). In models that assumed a constant time lag between the optimum laying date and the food peak, we found that this lag was about 33 days, which is consistent with (although slightly lower than) what is known of the biology of this population: egg laying takes about 9 days, incubation 12 days and hence

it is optimal to have nestlings close to fledging (which is at day 16-18) at the food peak (Visser et al., 1998, 2006; Reed et al., 2013b; Gienapp et al., 2013).

POSSIBLE EXTENSIONS

Here we focused for simplicity on selection on a individual trait. A multivariate statistical model of stabilizing selection on multiple traits (analogue to the multiple linear regression in Lande and Arnold (1983)) is straightforward to formulate in our framework. After transformation of the GLMM, the fitness function becomes a multivariate Gaussian, with orientation and shape determined by a symmetric matrix whose diagonal elements measure stabilizing selection, and off-diagonal elements measure correlational selection (Lande, 1979). In a stochastic environment, stationary fluctuations in the optimum would then also be multivariate, characterized by the covariance matrix of the stationary joint distribution of optima for all traits (rather than the single variance parameter used here), and a non-symmetric cross-covariance matrix representing the autocovariance over one time step (rather than the single autocorrelation parameter used here) (see e.g. appendix in Chevin (2013)). In practice however, the number of parameters to estimate is likely to become problematic. For k measured traits, the 3 matrices described above require estimation of $k(2k + 1)$ parameters in total. The slope of the k -dimensional optimum relative to n environmental variables (environmental sensitivity of selection) would involve kn additional parameters. Estimation of these parameters would require not only a large number of time points, but also a huge sample size at each time point.

We focused on a single selection episode, acting either through a single fitness component, or through an integrative measurement of lifetime fitness. A natural extension of our method would be to combine multiple selection episodes, acting through different fitness components. For instance in our study case, we could consider clutch failure (causing zero-inflation in the distribution of fledgling number) as a separate episode of selection, which may be correlated to some extent with the selection episode we focused on above, based on the nonzero-inflated part of the distribution. More generally, it would be useful to model selection operating through all possible vital rates (age-specific sur-

vivals and fecundities), in an age-structured population. Several approaches have been proposed to integrate different fitness components into measurements of selection, from combining selection gradients measured in each selection episode (Arnold and Wade, 1984), to aster modeling of lifetime reproductive success (Shaw and Geyer, 2010), or explicit age-structured demographical models (Lande, 1982; van Tienderen, 2000; Engen et al., 2012). Engen et al. (2011) showed that when (weak) stabilizing selection operates on a trait through all vital rates in an age-structured population, the resulting overall selection on this trait is well approximated by a single Gaussian function with width and optimum that are averages of those at each age, weighted by the relevant reproductive values. Engen et al. (2012) further showed how to measure fluctuating selection in an age-structured population, but assuming no autocorrelation, and focusing on linear gradients. All these approaches could be combined with the present method in order to measure autocorrelated fluctuations in age-specific optima, and a resulting single optimum for overall selection. A likely outcome will be that overall selection through lifetime fitness fluctuates less than selection at each episode (e.g. one breeding season), because of a buffering of environmental fluctuations as individuals experience different optima throughout their life. Note that a serious complication of such an approach would be that optimum phenotypes (and more generally selection strengths) are likely to be correlated across ages, especially if they depend on the same unobserved environmental variables.

Our method can be applied to analyze spatial (rather than temporal) variation in phenotypic selection, as long as there is no dispersal during selection. For one-dimensional space (e.g., latitudinal or altitudinal transects), the analyses conducted above are readily transposable, using spatial coordinates instead of time. The first-order autoregressive process should be replaced by its continuous equivalent (an Ornstein-Uhlenbeck process), where the autocorrelation over one step is replaced by a rate of exponential decay of autocorrelation with distance. A tendency with space would then indicate a spatial cline in the optimum, while the random process would estimate how the random component of phenotypic selection varies (and correlates) across space. Spatio-temporal variation in phenotypic selection would also be straightforward to analyze under the assumption

that temporal and spatial effects act additively on the optimum. Such a model could represent permanent differences between optimal egg-laying dates at different spatial locations, arising for instance from topography or nutrient availability. Beyond this, analysis of spatio-temporal variation in selection (including the spatial scale of synchrony in selection) would be more involved, and requires further developments.

Conclusion

We introduced an approach that connects empirical measurements of traits and fitness across time to theoretical predictions for evolution under environmentally driven fluctuating selection. Applications of this method and its future extensions to datasets from natural populations should provide a more quantitative picture of the eco-evolutionary consequences of randomly changing environments.

Acknowledgements

We thank R. G. Shaw, J. Hadfield, and two anonymous reviewers for constructive comments. L.-M. C. is supported by the grants ContempEvol (ANR-11-PDOC-005-01) and PEPS (ANR-12-ADAP-0006) from the Agence Nationale de la Recherche, and M. E. V is supported by the European Research Council (ERC-2013-AdG 339092).

Literature Cited

- Ahola, M. P., T. Laaksonen, T. Eeva, and E. Lehikoinen, 2012. Selection on laying date is connected to breeding density in the pied flycatcher. *Oecologia* 168:703–710.
- Arnold, S. J., M. E. Pfrender, and A. G. Jones, 2001. The adaptive landscape as a conceptual bridge between micro-and macroevolution. *Genetica* 112:9–32.
- Arnold, S. J. and M. J. Wade, 1984. On the measurement of natural and sexual selection: Theory. *Evolution* 38:709–719.

- Bell, G., 2010. Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Phil. Trans. Roy. Soc. B* 365:87–97.
- Brown, C., M. Brown, and E. Roche, 2013. Fluctuating viability selection on morphology of cliff swallows is driven by climate. *Journal of Evolutionary Biology* 26:1129–1142.
- Bull, J., 1987. Evolution of phenotypic variance. *Evolution* 41:303–315.
- Bulmer, M., 1974. Density-dependent selection and character displacement. *Am. Nat.* 108:45–58.
- Bürger, R. and A. Gimelfarb, 2002. Fluctuating Environments and the Role of Mutation in Maintaining Quantitative Genetic Variation. *Genetical research* 80:31–46.
- Burnham, K. P. and D. R. Anderson, 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer.
- Cain, A., L. Cook, and J. Currey, 1990. Population size and morph frequency in a long-term study of *cepaea nemoralis*. *Proc. Roy. Soc. Lond. B* 240:231–250.
- Calsbeek, R., T. P. Gosden, S. R. Kuchta, and E. Svensson, 2012. Fluctuating selection and dynamic adaptive landscapes. Pp. 89–103, *in* E. Svensson and R. Calsbeek, eds. *The adaptive landscape in evolutionary biology*. Oxford University Press.
- Charlesworth, B., 1993. Directional selection and the evolution of sex and recombination. *Genetical Research* 61:205–224.
- 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., 2013. Genetic constraints on adaptation to a changing environment. *Evolution* 67:708–721.
- 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. and R. Lande, 2010. When do adaptive plasticity and genetic evolution prevent extinction of a density-regulated population? *Evolution* 64:1143–50.
- Dieckmann, U. and M. Doebeli, 1999. On the origin of species by sympatric speciation. *Nature* 400:354–7.
- Engen, S., R. Lande, and B.-E. Sæther, 2011. Evolution of a plastic quantitative trait in an age-structured population in a fluctuating environment. *Evolution* 65:2893–2906.
- Engen, S. and B.-E. Sæther, 2014. Evolution in fluctuating environments: Decomposing selection into additive components of the robertson–price equation. *Evolution* 68:854–865.
- Engen, S., B.-E. Saether, T. Kvalnes, and H. Jensen, 2012. Estimating fluctuating selection in age-structured populations. *J. Evol. Biol.* 25:1487–1499.
- Estes, S. and S. J. Arnold, 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. *Am. Nat.* 169:227–44.
- Felsenstein, J., 1976. The theoretical population genetics of variable selection and migration. *Ann. Rev. Genet.* 10:253–280.
- Ferkingstad, E. and H. Rue, 2015. Improving the INLA approach for approximate Bayesian inference for latent Gaussian models Pp. 1–15. ArXiv preprint arXiv:1503.07307v3.
- Gavrilets, S. and S. M. Scheiner, 1993. The genetics of phenotypic of reaction norm shape V . Evolution of reaction norm shape. *J. Evol. Biol.* 6:31–48.
- Gelman, A., J. Hwang, and A. Vehtari, 2014. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 24:997–1016.
- 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? *Phil. Tran. Roy. Soc. B* 368:20120289.

- Gillespie, J. H., 1973. Natural selection with varying selection coefficients—a haploid model. *Genetical Research* 21:115–120.
- , 1974. Natural selection for within-generation variance in offspring number. *Genetics* 76:601–606.
- Gneiting, T. and A. E. Raftery, 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Stat. Assoc.* 102:359–378.
- Grant, P. R. and B. R. Grant, 2002. Unpredictable evolution in a 30-year study of darwin’s finches. *Science* 296:707–711.
- Hansen, T. F., J. Pienaar, and S. H. Orzack, 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* 62:1965–1977.
- Hedrick, P. W., 2006. Genetic polymorphism in heterogeneous environments: the age of genomics. *Ann. Rev. Ecol. Evol. Syst.* Pp. 67–93.
- Hedrick, P. W., M. E. Ginevan, and E. P. Ewing, 1976. Genetic polymorphism in heterogeneous environments. *Ann. Rev. Ecol. Evol. Syst.* Pp. 1–32.
- Held, L., B. Schrödle, and H. Rue, 2010. Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA. Pp. 91–110, *in* T. Kneib and L. Fahrmeir, eds. *Statistical Modelling and Regression Structures*. Physica-Verlag HD, Heidelberg. URL http://www.springerlink.com/index/10.1007/978-3-7908-2413-1_6.
- Hereford, J., 2009. A quantitative survey of local adaptation and fitness trade-offs. *Am. Nat.* 173:579–588.
- Holt, R. D., 2009. Bringing the hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc. Nat. Acad. Sci.* 106:19659–19665.
- Janzen, F. and H. Stern, 1998. Logistic regression for empirical studies of multivariate selection. *Evolution* 52:1564–1571.

- Kelly, C. A., 1992. Spatial and temporal variation in selection on correlated life-history traits and plant size in *chamaecrista fasciculata*. *Evolution* Pp. 1658–1673.
- Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. Hill, A. Hoang, P. Gibert, and P. Beerli, 2001. The strength of phenotypic selection in natural populations. *The American Naturalist* 157:245–261.
- Lande, R., 1975. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetical Research* 26:221–235.
- , 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* Pp. 314–334.
- , 1979. Quantitative Genetic Analysis of Multivariate Evolution , Applied to Brain : Body Size Allometry. *Evolution* 33:402–416.
- , 1982. A quantitative genetic theory of life history evolution. *Ecology* 63:607–615.
- , 2009. Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation. *J. Evol. Biol.* 22:1435–46.
- Lande, R. and S. J. Arnold, 1983. The measurement of selection on correlated characters. *Evolution* 37:1210–1226.
- Lande, R. and S. Shannon, 1996. The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution* 50:434–437.
- MacColl, A. D., 2011. The ecological causes of evolution. *Trends Ecol. Evol.* 26:514–522.
- Michel, M. J., L.-M. Chevin, and J. H. Knouft, 2014. Evolution of phenotype-environment associations by genetic responses to selection and phenotypic plasticity in a temporally autocorrelated environment. *Evolution* 68:1374–1384.
- Mills, S. and J. Beatty, 1979. The Propensity Interpretation of Fitness. *Philosophy of Science* 46:263–286.

- Mitchell-Olds, T. and R. G. Shaw, 1987. Regression analysis of natural selection: statistical inference and biological interpretation. *Evolution* Pp. 1149–1161.
- Morrissey, M. B. and J. D. Hadfield, 2012. Directional selection in temporally replicated studies is remarkably consistent. *Evolution* 66:435–42.
- Morrissey, M. B. and K. Sakrejda, 2013. Unification of regression-based methods for the analysis of natural selection. *Evolution* 67:2094–2100.
- Petris, G., S. Petrone, and P. Campagnoli, 2009. *Dynamic Linear Models with R*. Springer.
- Phillips, P. and S. Arnold, 1989a. Visualizing multivariate selection. *Evolution* 43:1209–1222.
- Phillips, P. C. and S. J. Arnold, 1989b. Visualizing multivariate selection. *Evolution* Pp. 1209–1222.
- 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. *J. Anim. Ecol.* 82:131–144.
- Reimchen, T. and P. Nosil, 2002. Temporal variation in divergent selection on spine number in threespine stickleback. *Evolution* 56:2472–2483.
- Rice, S. H., 2004. *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates, Inc.
- Rue, H., S. Martino, and N. Chopin, 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71:319–392.

- Schluter, D., 1988. Estimating the form of natural selection on a quantitative trait. *Evolution* Pp. 849–861.
- Shaw, R. G. and C. J. Geyer, 2010. Inferring fitness landscapes. *Evolution* 64:2510–2520.
- Shumway, R. H. and D. S. Stoffer, 2010. Time series analysis and its applications: with R examples. Springer, New York, NY.
- Siepielski, A. M., J. D. DiBattista, and S. M. Carlson, 2009. It’s about time: the temporal dynamics of phenotypic selection in the wild. *Ecology letters* 12:1261–76.
- Siepielski, A. M., K. M. Gotanda, M. B. Morrissey, S. E. Diamond, J. D. DiBattista, and S. M. Carlson, 2013. The spatial patterns of directional phenotypic selection. *Ecology Letters* 16:1382–92.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde, 2002. Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc. B* 64:583–639.
- Svardal, H., C. Rueffler, and J. Hermisson, 2011. Comparing environmental and genetic variance as adaptive response to fluctuating selection. *Evolution* 65:2492–513.
- Svensson, E. and R. Calsbeek, 2012. The adaptive landscape in evolutionary biology. Oxford University Press, Oxford, UK.
- van Tienderen, P. H., 2000. Elasticities and the link between demographic and evolutionary dynamics. *Ecology* 81:666–679.
- Tufto, J., 2015. Genetic evolution, plasticity and bet-hedging as adaptive responses to temporally autocorrelated fluctuating selection: A quantitative genetic model. *Evolution*, in press.
- Vedder, O., S. Bouwhuis, and B. C. Sheldon, 2013. Quantitative assessment of the importance of phenotypic plasticity in adaptation to climate change in wild bird populations. *PLoS biology* 11:e1001605.

- Via, S. and R. Lande, 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution* 39:505–522.
- Visser, M., A. Van Noordwijk, J. Tinbergen, and C. Lessells, 1998. Warmer springs lead to mistimed reproduction in great tits (*parus major*). *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265:1867–1870.
- Visser, M. E., P. Gienapp, A. Husby, M. Morrissey, I. de la Hera, F. Pulido, and C. Both, 2015. Effects of spring temperatures on the strength of selection on timing of reproduction in a long-distance migratory bird. *PLoS Biology* 13:e1002120.
- 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.
- Visser, M. E. and J. J. Sanz, 2009. Solar activity affects avian timing of reproduction. *Biology Letters* 5:739–742.
- Wade, M. J. and S. Kalisz, 1990. The causes of natural selection. *Evolution* 44:1947–1955.
- Zhang, X.-S., 2012. Fisher’s geometrical model of fitness landscape and variance in fitness within a changing environment. *Evolution* 66:2350–2368.